Statistical Approaches for Next-Generation Sequencing Data

Statistical Approaches for Next-Generation Sequencing Data

A dissertation presented

by

Dandi Qiao

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 2012

Dissertation Advisor: Professor Christoph Lange Dandi Qiao

Statistical Approaches for Next-Generation Sequencing Data

Abstract

During the last two decades, genotyping technology has advanced rapidly, which enabled the tremendous success of genome-wide association studies (GWAS) in the search of disease susceptibility loci (DSLs). However, only a small fraction of the overall predicted heritability can be explained by the DSLs discovered. One possible explanation for this "missing heritability" phenomenon is that many causal variants are rare. The recent development of high-throughput next-generation sequencing (NGS) technology provides the instrument to look closely at these rare variants with precision and efficiency. However, new approaches for both the storage and analysis of sequencing data are in imminent needs.

In this thesis, we introduce three methods that could be utilized in the management and analysis of sequencing data. In Chapter 1, we propose a novel and simple algorithm for compressing sequencing data that leverages on the scarcity of rare variant data, which enables the storage and analysis of sequencing data efficiently in current hardware environment. We also provide a C++ implementation that supports direct and parallel loading of the compressed format without requiring extra time for decompression.

Chapter 2 and 3 focus on the association analysis of sequencing data in population-based design. In Chapter 2, we present a statistical methodology that allows the identification of genetic outliers to obtain a genetically homogeneous subpopulation, which reduces the false positives due to population substructure. Our approach is computationally efficient that can be applied to all the genetic loci in the data and does not require pruning of variants in linkage disequilibrium (LD). In Chapter 3, we propose a general analysis framework in which thousands of genetic loci can be tested simultaneously for association with complex phenotypes. The approach is built on spatial-clustering

methodology, assuming that genetic loci that are associated with the target phenotype cluster in certain genomic regions. In contrast to standard methodology for multi-loci analysis, which has focused on the dimension reduction of data, the proposed approach profits from the availability of large numbers of genetic loci. Thus it will be especially relevant for whole-genome sequencing studies which commonly record several thousand loci per gene.

Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to express my very great appreciation to my advisor Christoph Lange, who has helped and guided me through my doctoral study. He has always been supportive and patient, and I feel very lucky to have him as my advisor. I also would like to thank my committee members, Xihong Lin and Scott T. Weiss, who have given me many insightful suggestions and ideas, and are always available for me even when they have a completely full schedule. I am also grateful for the comments and advices from Nan Laird.

Secondly, I would like to thank Wai-Ki Yip for his time and efforts for providing us with a stable and fast computing environment, and his help with the datasets used in my research. My grateful thanks are also extended to Kaustubh Adhikari, who had spent hours discussing with me about genetics and introduced me to useful approaches. I also appreciate the comments and suggestions from Manuel Mattheisen for my research, and the help from Michael Cho, Ed Silverman and Jin Zhou on the COPDgene data.

Thirdly, I really appreciate the support and help from my friends, who have inspired me with their experiences and positive attitudes, including Xinyi Lin, Yifan Zhang, Jun Li, Linda Valeri, Wei Dai, Tamar Tsivion, Danielle Braun and Xuefeng Wang. I am also grateful to have met so many great people in the department.

Most of all, my special thanks are to my mom, dad, grandma and my fiance James, who have given me unlimited support and courage, who have shown me the true meaning of a family. I am so lucky to always have you at my back and I know no matter where I am or how old I am, as long as you are there, that place is called home.

Handling the data management needs of high-throughput sequencing data: SpeedGene, a compression algorithm for the efficient storage of genetic data

Dandi Qiao, Wai-Ki Yip, Christoph Lange

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

1.1 Introduction

As the influx of high-throughput sequencing data [The 1000 Genome Project Consortium, 2010], [Bansal et al., 2010], [Metzker, 2010] is imminent, the data management requirements for the analysis packages have changed fundamentally. While, during the days of candidate gene analysis and linkage analysis, ”only” up to several thousands of genetic loci had to be stored and loaded into the analysis packages, current Genome-wide Association studies (GWAS) provide genetic information on several millions of genetic loci. Thus, the typical size of a dataset containing mostly common variants is about 1 to 30 Gigabytes. For high-throughput sequencing studies, the number of genetic loci genotyped increases by several magnitudes, and the file size of such sequencing data can be up to several Terabytes. For such large files, the loading process can take up to few hours without counting the time for analysis. This results in great waste of disk space and computation time, which is a problem that is encountered routinely.

One possible solution is to use the general-purpose compression software, such as Gzip and BGZip. However, such compression software is not designed specifically for genetic data and its analysis, so the compression rate is relatively low and decompression is always needed before accessing the data. Better solutions have been proposed. PLINK and PBAT, which are free whole-genome association analysis toolsets, have introduced Binary PED formats [Lange et al., 2004] [Purcell et al., 2007]. This format ensures that only 2 Bits are required for storing the information of one genotype. It is the most popular compression format used in GWAS. However, the compression rate is not sufficient for massive datasets generated nowadays as their compressed datasets could still occupy several Gigabytes of the disk space. In recent years, sophisticated compression techniques designed specifically for sequencing data have been proposed. For example, DNAzip [Christley et al., 2009] introduced the idea of storing only the difference between one individual genome data and a reference genome. However, such algorithms suffer the large overhead for storing the reference genome. Also, they require substantial CPU-time for decompression.

We propose here a simple and efficient algorithm to store large datasets containing SNP data of multiple samples. We show that our algorithm always works better than the compression algorithm implemented in PLINK or PBAT and provides excellent compression rate for sequencing data. Also, the compressed data structure provides the potential for efficient implementation of permutation methods and does not require any overhead CPU-time for decompression. We have implemented the algorithm in the GPL licensed C++ library: SpeedGene. We show that it takes much less time for loading the compressed files than PLINK using our library. In addition, Our C++ implementation supports parallel loading of the genetic information, which further decreases the loading time as the number of parallel jobs increases. The version 1.0 of the SpeedGene library is available at `http://people.hsph.harvard.edu/~dqiao/SpeedGene.html` together with detailed instructions and examples.

1.2 Methods

1.2.1 The LINKAGE/PLINK data format

The LINKAGE or PLINK data format is a commonly used data format for storing SNP data in Genome-Wide Association studies. Data files in this format is called pedigree files and have ".ped" as the suffix. This format can be converted from or to the VCF format used in 1000 Genome Project using VCFtools [Danecek et al., 2011]. The SpeedGene library currently only recognizes pedigree files in the LINKAGE/PLINK format, but the algorithm can be implemented for compressing SNP data in the VCF format. The VCF format requires the same amount of disk space for each genotype (4 Bytes) as the LINKAGE/PLINK format, so the compression rate of this algorithm applying on VCF files should be similar to the compression rate for pedigree files. Note that VCF files may contain other informations such as Indels, Deletions, and the phase information, which could not be incorporated into the LINKAGE format. However, since SNP data are very commonly used genetic data in association studies and takes the most disk space,

efficient storage of the SNP data could still save a lot of resources. In the demonstration of the algorithm and the examples below, we use the LINKAGE/PLINK format as the input format.

```
A1 A2 A3 A4 A5 A6
1 1 0 0 1 2 1 2 3 3 2 4 3 1 0 0 2 2
1 8 0 0 2 1 2 2 2 3 0 0 3 1 2 2 2 3
1 3 1 2 1 2 1 2 3 2 2 2 3 3 1 1 2 3
1 4 0 0 2 1 1 0 3 2 4 4 3 3 1 1 2 2
1 5 0 0 1 2 1 2 3 3 2 4 3 1 0 0 2 2
1 6 0 0 2 1 2 2 2 3 0 0 3 1 2 2 2 3
```

Figure 1.1: A toy example of a pedigree file in the LINKAGE format. The first line contains the marker names. Starting from the second line, each line contains the pedigree and genetic information for each individual. The first six columns indicate the subject's pedigree ID, subject ID, father ID, mother ID, sex and affection status. The other columns contain the genetic data.

Any pedigree file in the LINKAGE format has the same structure, a toy example is shown in Figure 1.1. The first line contains the marker names, separated by a space character. Starting from the second line, each line includes pedigree and genetic information for each individual. The first six columns of these lines specify each individual's pedigree information in the order of pedigree ID, subject ID, father ID, mother ID, sex, and affection status. Subject ID must be unique within one's family. Father and mother ID could be 0 if this information is unknown, e.g. population-based study of unrelated subjects. Sex is 1 for male and 2 for female. Affection status is 1 if the subject is unaffected, 2 if affected, and 0 if the status is unknown. The other columns contain the genetic data for

each individual, separated by a space between each marker. Two columns are required to represent the information for two alleles, separated by a space. The allele information is coded using 0 to 4 where $1 = A, 2 = C, 3 = G, 4 = T$ and 0 represents missing allele information.

1.2.2 The SpeedGene Algorithm

The SpeedGene algorithm consists of three different sub-algorithms, which are selected by SpeedGene based on the minor allele frequency (MAF) of the genetic locus to be stored. The space needed for the compressed data is computed for the sub-algorithms beforehand. The SpeedGene algorithm then selects the best procedure among the three compression methods. The first sub-algorithm is based on the binary format implemented in PLINK and PBAT. It utilizes the fact that the marker information of each marker can be represented using a 2-digit binary number. The second sub-algorithm uses subject indices to indicate heterogeneous, homogeneous and missing genotypes. The third sub-algorithm uses binary digits to indicate heterogeneous genotype and subject indices to indicate homozygous and missing genotypes. A feature of all three compression methods is that the required memory space for storage can be computed prior to compression. Thereby, the SpeedGene algorithm is able to select the optimal method before compressing the data. The three sub-algorithms are described in detail in the following sections.

Sub-algorithm I: Compression using binary encoding

For any pedigree file, we assume that there are only bi-allelic markers in the file. For any allele of a marker, an individual may only have 0, 1 or 2 of this allele. Also, the allele information can be missing for any individual at any marker. Thus, the marker information can be transformed into the number of copies of a particular allele. It could be 0,1,2, or missing and could be converted to a 2-digit binary number. In the compression process, we find the minor alleles at each marker and use 00, 01, 10 to represent zero, one or two copies of the minor allele at one marker. 11 indicate that the genetic information is missing at this marker for the individual. Thus, one genotype in

the original file can be converted into two binary digits, which is 2 Bits on disk space. Four of such 2-digit binary number is 8 Bits, which equals 1 Byte. Therefore, the genetic information of four markers for one individual can be converted into 1 Byte in a binary file. This binary encoding is similar to the binary format used in PLINK [Purcell et al., 2007] or PBAT [Lange et al., 2004].

Based on this conversion method, we can compress the genetic information in the pedigree file into a much smaller binary file. As we have seen in the example (Figure 1.2), the genetic information for four genotypes occupies 16 Bytes in the original pedigree file, and it is converted to only 1 Byte in the compressed file, which could save up to a factor of sixteen on the disk space. If there are n subjects in the dataset, the storage requirement for compressing n genotypes for one marker using this algorithm is given by

$$\lceil 2 * n/8 \rceil \quad Bytes \tag{1.1}$$

For the assessment of the performance of the proposed SpeedGene algorithm, we will use the LINKAGE/PLINK format and the binary-encoding algorithm described above as the standard approach to which the SpeedGene algorithm will be compared.

Sub-algorithm II: Compression using subject indices

With the binary-encoding algorithm described above, the genetic information of any marker in one dataset is compressed to the same size since the compression algorithm does not depend on the frequency of each genotype. As we will see later, the performance of the binary compression is the best we can achieve when the variants are relatively common ($MAF > 30\%$). However, for SNPs with small MAF, only a few subjects have the heterozygous genotype and, even fewer, have the rare homozygous genotype. Thus, it is wasting disk space if the genetic information for all the subjects is recorded, especially for the subjects with the common homozygous genotypes which is by far the most frequent genotype. Therefore, we can utilize this feature of SNPs with small MAF, and record only the indices of the subjects with the missing, heterozygous or rare homozygous genotypes for the SNP. The common homozygous genotype is the default

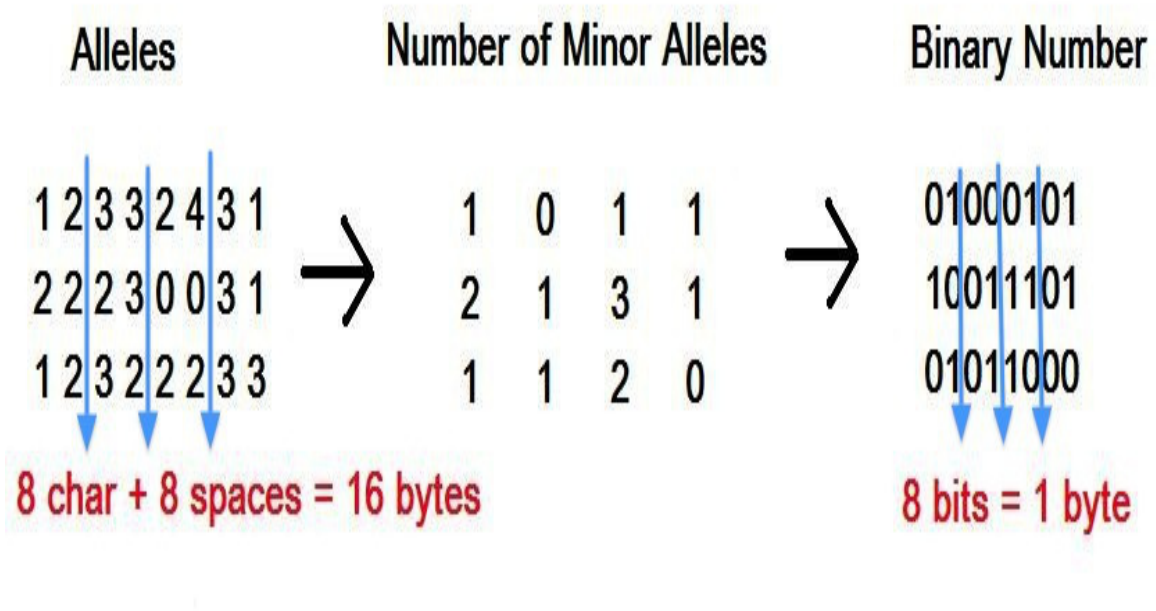

Figure 1.2: Genetic information of the first four markers for the first three individuals in the toy example is extracted here to demonstrate the sub-algorithm I. Each row represents the four genotypes of one individual.The minor alleles for the four markers are assumed to be 2 2 2 1 respectively, and are underlined. Genotype 0 0 represents missing genotypes in the original dataset, which is converted to 3 to indicate missing genotypes.

genotype. Since most of the SNPs of the human genome have small MAF [The 1000 Genome Project Consortium, 2010], the improvements of this approach is substantial compared to the binary-encoding algorithm in the last section.

Specifically, suppose we have n subjects in the data, then we need $\lceil log_2(n) \rceil$ binary digits in order to record the index of any subject. First, the number of the rare homozygous, the heterozygous and the missing genotypes are counted. This information is used to calculate the compressed size and determine whether Sub-algorithm II should be used for the SNP. If Sub-algorithm II requires the smallest amount of memory, SpeedGene will use Sub-algorithm II for the compression of the genetic data for the SNP. The indices of the subjects with the homozygous, heterozygous and missing genotypes are transformed into binary digits and are written into the binary file afterwards. Since the number of subjects with each genotype varies, the counts, each requires $\lceil log_2(n) \rceil$ Bits on the disk space, are written to the file before the indices of the subjects are outputted to the file. Thus, the storage requirement for compressing n genotypes for one marker using this algorithm is given by

$$\lceil \lceil log_2(n) \rceil * (\#Homo + 1 + \#Heter + 1 + \#Missing + 1)/8 \rceil \quad Bytes \tag{1.2}$$

where #Missing denotes the number of subjects with the missing genotype, #Homo denotes the number of subjects with the rare homozygous genotype, and #Heter denotes the number of subjects with the heterozygous genotype.

Sub-algorithm III: Compression using binary encoding and subject indices

As we will see in the next section, Sub-algorithm II works best for SNPs with very small MAF, but performs worse than Sub-algorithm I for more common SNPs ($MAF > 0.3$). However, by combining Sub-algorithm I and II, we can create a hybrid approach that performs better than Sub-algorithm I and II for SNPs whose MAFs are somewhere between uncommon and very common.

Since the heterozygous genotype is more common for genetic loci that are in the range between uncommon and very common ($0.05 \leq MAF \leq 0.3$), recording the heterozygous genotype by the indices of subjects is not very efficient. Instead we use a binary number of n digits to indicate the subjects with the heterozygous genotype, where n is the number of subjects in the dataset. If subject i has the heterozygous genotype for the SNP, 1 is put at position i instead of 0. Beside this, the indices of subjects with the missing and homozygous minor allele genotypes are recorded in the same way as in Sub-algorithm II. The storage requirement of the marker information for n samples using this algorithm is given by

$$\lceil((\#Homo + 1 + \#Missing + 1) * \lceil log_2(n) \rceil + n)/8 \rceil \quad Bytes \tag{1.3}$$

where #Homo denotes the number of subjects with the rare homozygous genotype and #Missing denotes the number of subjects with the missing genotype for the SNP.

For Sub-algorithm II and III, since the indices of the heterozygous and homozygous genotypes are stored for each marker, this compressed data structure makes computation for permutation methods much convenient.

1.3 Results

1.3.1 Performance Comparison of Sub-algorithms

The SpeedGene algorithm selects for each genetic locus the optimal algorithm in terms of storage space (1.1-1.3) among the three sub-algorithms described above. To assess the performance of the SpeedGene algorithm, we compare it with the standard LINKAGE/PLINK format and the PLINK/PBAT compression algorithm. The efficiency of the SpeedGene algorithm depends on two factors, the genotype frequency of the genetic locus and the number of subjects included in the dataset. Assuming Hardy-Weinberg equilibrium, the first plot of Figure 1.3 gives a plot of the compression factor of the three sub-algorithms versus different MAFs for a dataset of 1000 subjects. The second plot shows the number of Bits needed per genotype for storing the genotype information of 1000 subjects at different MAF values. The dashed line provides the performance for the

SpeedGene algorithm which is based on the allele frequency and formulas 1.1, 1.2 and 1.3 to select the optimal compression procedure among Sub-algorithm I-III.

As in the plot, approximately, SpeedGene always achieves a compression factor of 16 compared to the standard LINKAGE format for $MAF > 0.3$ for which Sub-algorithm I is used. SpeedGene accomplishes a compression factor of 16 up to 30 compared to the LINKAGE/PLINK format for $0.05 \leq MAF \leq 0.3$ for which Sub-algorithm II is selected. For rare and uncommon alleles ($MAF < 0.05$), a compression factor of at least 30 compared to the LINKAGE format is realized. With smaller MAFs, the compression factor increases rapidly. Equivalently, 2 Bits per genotype would be needed for $MAF > 0.3$, about 1.0 to 2.0 Bits per genotype for $0.05 \leq MAF \leq 0.3$, and less than 1 Bit per genotype is needed for $MAF < 0.05$.

The performance of the algorithms also depends on the number of subjects in the dataset. Figure 1.4 shows the compression factor of the algorithms for one marker for different number of subjects, at eight MAF levels.
Generally, the compression factor decreases slightly as the number of subjects included increases, but is mostly constant over the range of number of subjects we have considered for different values of MAF.
In addition to that, the plots give us similar information as the plots above. For example, for SNP with $MAF = 0.01$, Sub-algorithm II is able to compress the genetic information by a factor of at least 100, which is much better than Sub-algorithm I and III. Thus, MAF is the most influential factor in determining which algorithm is the optimal method among the three sub-algorithms.

1.3.2 The C++ Library Implementation

We have implemented the algorithm in a C++ library called SpeedGene. There are two classes in the SpeedGene library. The first one is the Comp class, which is responsible for compressing a pedigree file in the LINKAGE/PLINK format into a text file that contains the subject information and a binary file that contains the genetic information. The binary

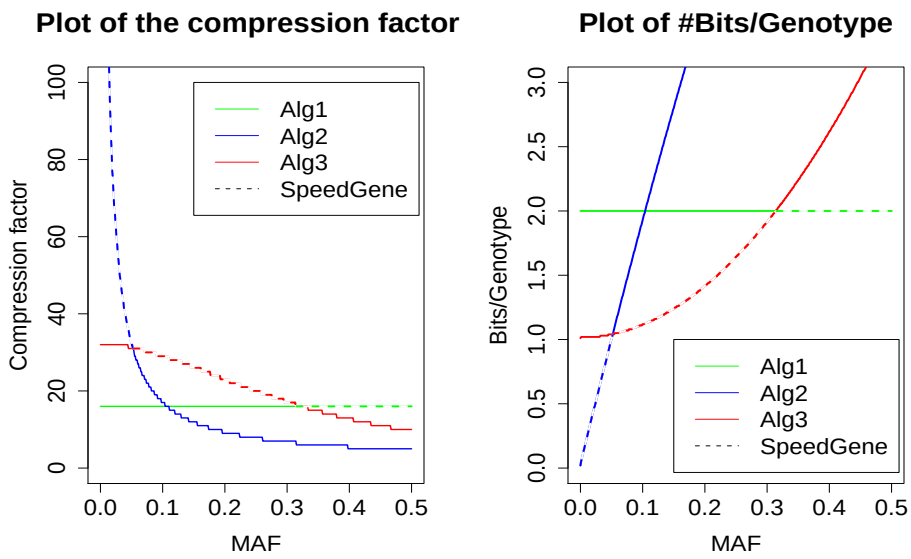

Figure 1.3: In the left plot, the compression factors of the three sub-algorithms are plotted against different MAF levels. In the right plot, the number of Bits needed for storing one genotype is plotted against different MAF levels. 1000 genotypes are simulated for one SNP at each MAF level. The space needed to store this information for one SNP in the LINKAGE/PLINK format is 4000 Bytes. The compression factor is the number of times by which the compressed file is smaller than the original file size (4000 Bytes).

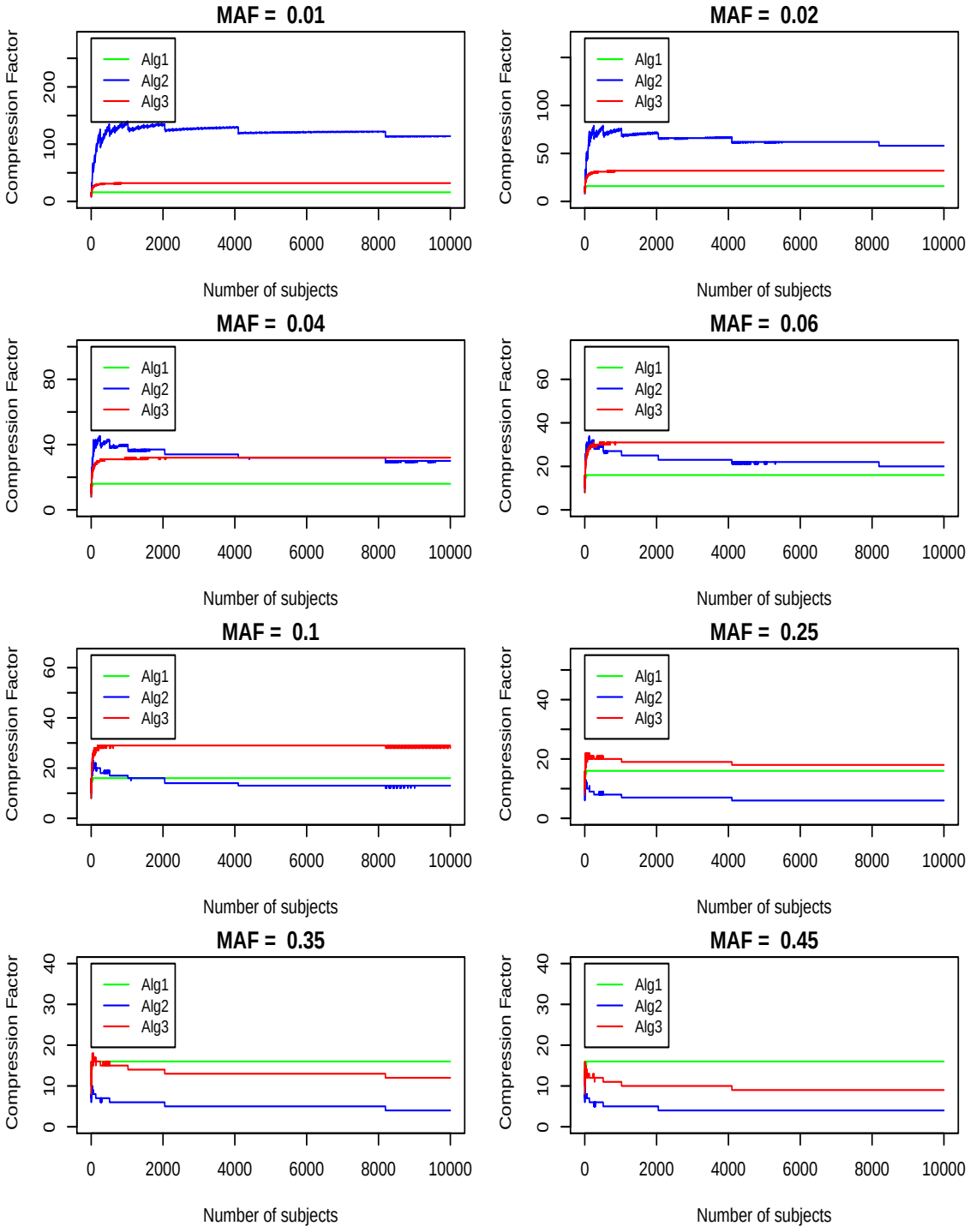

Figure 1.4: The compression factors of the three sub-algorithms are plotted against the number of subjects included in one dataset at eight different MAF levels, which are 0.01, 0.02, 0.04, 0.06, 0.1, 0.25, 0.35 and 0.45. The datasets we considered include at least 100 subjects and contain only one marker with the specified MAF level.

file is not human-readable and can only be used by the second class in our library. The compression process requires two scans of the pedigree file to avoid storing all the marker information before compression, which would take a great amount of memory space. The second class is the LoadComp class. As its name suggested, it is responsible for loading the compressed files into the memory, and for processing queries from the user. It provides an option to load the entire pedigree file or to load a section of the file. This partial-loading function ensures that only necessary information is loaded for the jobs that are running in parallel, which greatly decreases the loading time. Moreover, the public functions provided by the library allow the user to retrieve any information stored in the original file. This C++ library makes it straightforward for users to incorporate it into their own programs whereas other existing libraries do not offer such capability.

1.3.3 Performance

Compression rate

We evaluated the performance of the SpeedGene algorithm on two rare variants datasets. We simulated two datasets with 1000 subjects from the Wright's distribution [Wright, 1949], which is $f(p) = cp^{\beta_s - 1}(1-p)^{\beta_n - 1}e^{\sigma(1-p)}$, where the scaled mutation rates $\beta_s = 0.001, \beta_n = \beta_s/3$, the selection rate $\sigma = 12$, and c is a normalizing constant. Table 1.1 below shows the compressed file size for the simulated data. For sequencing data, the optimal algorithm is Sub-algorithm II for most of the SNPs. Thus, SpeedGene is able to achieve a large compression rate. In the simulated data, the compression factor is approximately 200, which is equivalent to 0.16 Bits per genotype, whereas 2 Bits per genotype is required by PLINK or PBAT. Gzip seems to perform much better on rare variant data than on common variant data, however, such general-purpose software takes extra time to decompress the files before loading them into the memory. We have also extrapolated the approximate file size if DNAzip is used [Christley et al., 2009]. According to the paper, each SNP for one person requires slightly less than 1 Byte per SNP for storage and it requires a reference human genome (~ 3 Gigabytes) and a reference SNP map (~ 1.2 Gigabytes) to retrieve the entire genome data.

Table 1.1: Compressed file sizes of the simulated datasets using PLINK, Gzip, SpeedGene and DNAzip. Each dataset contains 1000 subjects.

#SNPs	Size	PLINK	Gzip	SpeedGene	DNAzip	Avg MAF
1 million	3.731 GB	238 MB	22 MB	18 MB	16 MB+ $\sim$ 4.2 GB	0.004944
30 million	112 GB	6.985 GB	592 MB	534 MB	310 MB + $\sim$ 4.2 GB	0.004228

We also applied these methods to two real datasets. One dataset contains the genotype data from the Framingham Heart Study (FHS), which includes 6956 subjects and 340,444 SNPs. The other dataset is obtained from the COPDgene study on patients with Chronic Obstructive Pulmonary Disease (COPD). It includes 257 subjects with 162757 SNPs over the human genome and 77% of the SNPs in this sequencing data have a MAF $\leq 5\%$. The original file size and the compressed file sizes using different compression methods are shown in Table 1.2. For the FHS dataset, since that most of the SNPs are common, the compression rate of SpeedGene is just slightly greater than that of PLINK. Gzip gives a much lower compression ratio of 6 here, as for most common variant datasets. The COPDgene sequence data contains mostly rare variants, but still includes some common variants, so we observe a much higher compression rate with SpeedGene than with PLINK and Gzip.

Table 1.2: File sizes of the FHS dataset and COPDgene dataset, compressed using PLINK, SpeedGene and Gzip.

Dataset	Size	PLINK	Gzip	SpeedGene	Avg MAF
FHS	8.822 GB	564.6 MB	1.400 GB	460 MB	0.238637
COPDgene	161 MB	10.1 MB	20.5 MB	3.6 MB	0.057327

Loading time

The time for loading the compressed datasets using SpeedGene and PLINK on a 2.35GHz AMD Opteron CPU with 128GB of RAM is shown in Table 1.3 below. The time to load the entire file using SpeedGene is less than half of the time needed by PLINK for the simulated datasets. If the analysis is ran in parallel, the loading time using SpeedGene is decreased further as the number of jobs ran in parallel increases. For example, if we

are loading $1/10$ of the dataset with 30 million SNPs in each parallel job, the loading time needed by SpeedGene is 1.8 minute.

Table 1.3: The CPU time needed for loading the two compressed files using SpeedGene and PLINK on a 2.35GHz AMD Opteron CPU with 128GB of RAM.

Number of SNPs	Loading time (SpeedGene)	Loading time (PLINK)
1 million	26 sec	56 sec
30 million	11 min	29 min

1.4 Discussion

To tackle the problem of large file sizes and long loading times of genetic data, we have developed a new compression algorithm - SpeedGene. The algorithm selects the optimal approach among three methods in terms of the required disk space. We have shown that the algorithm always works better than the compression algorithms provided by PBAT and PLINK, and can reach a compression factor of sixteen up to few hundreds. Especially for sequencing data with mostly rare variants, the algorithm is able to compress files of hundreds of Gigabyte to hundreds of Megabytes. Similar compression rate can be reached for the VCF files containing SNP data. In addition, the compressed data structure requires no extra time for decompression and could reduce a large amount of computation time for performing permutations on the genotypes.

A C++ implementation of the SpeedGene algorithm is provided and an integration in R is ongoing, but the algorithm could be implemented easily for other data formats and using other programming languages. The SpeedGene library utilizes the structure of the compressed data and enables direct loading of the genotype data into the memory. Moreover, the functions in the LoadComp class of this library allow the user to flexibly retrieve any specified subject or genetic information from the compressed dataset. Furthermore, user-friendly parallel-loading function is supported, which in result shortens the loading time greatly when parallel jobs are dispatched in clusters.

To fully utilize the compression algorithm, it needs to be incorporated into other analysis software for association studies, where the genetic information can be loaded using the library and directly sent for analysis in the software. For example, we are planning to include this binary format as one of the standard input format in NPBAT, which is an interactive software for the analysis of population based genetic association studies. Such incorporation would require additional efforts, but with the gain of much more disk space and shorter loading time, it will be beneficial in the long run.

On association analysis of rare variants under population-substructure: An approach for the detection of subjects that can cause bias in the analysis

Dandi Qiao, Manuel Mattheisen, Christoph Lange

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA

2.1 Introduction

Genetic association analysis has proven to be a powerful statistical tool for the identification of disease loci in the human genome [Burton et al., 2007] [McCarthy et al., 2008] [Stranger et al., 2011]. Population-based association analysis is straightforward and computationally fast, even at a whole-genome level. One of the main caveats of population based association analysis, however, is that it can be susceptible to bias due to genetic confounding, i.e. population substructure.

This issue has been the focus of statistical research for some time. In designs of unrelated individuals, most genetic association tests take the form of a score test in which the numerator sums the contributions of the study subjects to the statistics and the denominator calculates the variance of the statistic, assuming independence of the study subjects. In the presence of mating among relatives or population substructure, the genotypes of the study subjects are no longer independent, leading to a potentially biased estimate for the variance of the test statistic. This can cause the test statistic to become anti-conservative. Genomic control approach adjusts for the bias in the variance of the test statistic by estimating a variance inflation factor at a set of reference loci and scaling the variance of the test statistic accordingly [Devlin and Roeder, 1999] [Reich et al., 2001b]. Recently, with the arrival of GWAS data, Principal Component Analysis gained popularity [Price et al., 2006] [Patterson et al., 2006]. They infer population substructure and admixture based on the Principal Component Analysis of the variance-covariance matrix of genotyped markers [McVean, 2009] [Novembre and Stephens, 2008]. Then, the principal components are either used to identify genetically homogenous subpopulations in the study [Luca et al., 2008] or to adjust the association for genetic confounding [Price et al., 2006]

For the association analysis of rare variants, the application of such approaches to avoid bias due to population substructure and admixture can be problematic. In PCA approach, the estimation of the variance/covariance matrix can become unstable for genetic loci with low minor allele frequencies, making the results of this approach less reliable. For

example, the investigators usually select markers with allele frequencies greater than 10 % before applying principal component analysis [He et al., 2011] [Sladek et al., 2007]. An alternative that could be considered here is to assess population substructure for loci with common alleles and apply the principal component results to the rare-variant analysis, assuming that the population substructures for rare and common variants are the same. The transferability of population substructure between common and rare genetic loci is a hypothesis which has not been assessed thoroughly based on real data so far. The general applicability of this concept seems to be problematic in light of the age of the different variant types, i.e. common variants are genetically much older than rare variants [Mathieson and McVean, 2012]. Although rare variant approaches rely mostly on permutation tests for the assessment of the significance, the concept of genomic control generally can be modified and applied to rare variant analysis. However, it can give a reduced power and cannot be utilized to identify homogeneous subpopulations.

Here, we proposed a simple, computationally fast approach that allows the identification of genetic outliers to obtain a genetically homogeneous subpopulation in studies with sequence data, minimizing the impact of population substructure on rare variants analysis. The approach is able to utilize the information on all available genetic loci and does not require the selection of a subset of markers that are not in linkage disequilibrium (LD). The test statistic is computed for each individual based on all the rare variant information available. The power and the type I error of the approach are examined in simulation studies and by the applications to the HapMap 3 and the 1000 Genome Project data. We compare the performance of our approach with the outlier detection algorithm based on PCA.

2.2 Material and Methods

2.2.1 Introducing test statistics T_1 and T_2

Suppose in a genetic association study of unrelated individuals, genotype data is available at m bi-allelic loci for all the study subjects. We denote the number of the minor alleles at the ith marker locus by X_i for one subject. We define the genetic residual by $\Delta X_i = X_i - E(X_i)$ where $E(X_i)$ is the expected number of the minor alleles at the ith locus in the study population. The genetic residual can be considered as the genetic deviation of the subject at i^{th} locus from the study population. We define two genome-wide scores that measure the distance between a particular individual and the population across the genome. The scores are given by

$$S_1 = \sum_{i=1}^{m} \Delta X_i = \sum_{i=1}^{m} (X_i - E(X_i))$$

and

$$S_2 = \sum_{i=1}^{m} |\Delta X_i| = \sum_{i=1}^{m} |X_i - E(X_i)|.$$

Based on the scores, we can construct the score tests T_1 and T_2 which are given by

$$T_1 = R_1^2 = \frac{(S_1 - E(S_1))^2}{Var(S_1)}$$

and

$$T_2 = R_2^2 = \frac{(S_2 - E(S_2))^2}{Var(S_2)}.$$

The first score aggregates the residuals over all the marker loci for one subject. If, for the study population and the population where the outliers are from, there is preferentially a one-direction difference in the MAF, i.e. most of the markers have smaller MAF in one population than in the other population, then the test score S_1 will be more powerful in detecting the population outliers. This situation can occur due to the founder effects in one subpopulation [Roy-Gagnon et al., 2011] [Reich et al., 2001a], long-range haplotypes [Price et al., 2008], etc. However, if the differences in minor allele frequencies between two subpopulations do not follow this patterns, test statistic S_2 is generally better suited to identify genetically different subjects. In Appendix A.3, we provide the

theoretically justification for that. We will further outline these features of the score statistics S_1 and S_2 in the simulation section of this paper.

Under the assumption of Hardy-Weinberg equilibrium, the expected marker score can be calculated based on the minor allele frequency, i.e. $E(X_i) = 2p_i$, where p_i is the true minor allele frequency at ith marker locus. For any real dataset, we can estimate the allele frequency p_i by the observed frequency of the minor allele in the actual data. Alternatively, the allele frequencies can be obtained from the corresponding reference populations. Assuming the absence of LD between the loci, the mean and variance of S_1 and S_2 can be derived analytically based on the allele frequencies, as shown in Appendix A.1. Then the test statistics are given by:

$$T_1 = \frac{[\sum_{i=1}^{m}(X_i - 2p_i])^2}{\sum_{i=1}^{m}(2p_i(1-p_i))} \tag{2.1}$$

$$T_2 = \frac{[\sum_{i=1}^{m}|X_i - E(X_i)| - (2p_i(1-p_i))]^2}{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))} \tag{2.2}$$

Then, under the null hypothesis of no population substructure, both test statistics T_1 and T_2 follow a χ^2 distribution with one degree of freedom.

2.2.2 Adjusting T_1 and T_2 in the presence of LD

For sequence data, the LD assumption may not always be reasonable unless only a subset of loci that are not in LD is selected. In the presence of LD, both standardized scores have to be adjusted accordingly. Since the variances of S_1 and S_2 do not depend on the actual genotype of the study subject and are constant across the subjects, ideally, we would need to adjust T_1 by

$$\frac{Var(\sum_{i=1}^{m}\Delta X_i)}{\sum_{i=1}^{m}Var(\Delta X_i)}$$

However, since the calculation of the correlations of the residuals across the genome requires a great amount of computation time, a genomic inflation factor for each test statistic can be estimated based on the distribution of the test statistic across the study subjects. For test statistic T_1, we estimate the genomic inflation λ_1 by

$$\hat{\lambda}_1 = \frac{\text{Median of } T_1 \text{ across all subjects}}{0.455} \tag{2.3}$$

where 0.455 is the 50th percentile of a $\chi^2_{(1)}$ distribution. Similarly for T_2, we estimate the genomic inflation factor λ_2 by

$$\hat{\lambda}_2 = \frac{\text{Median of } T_2 \text{ across all subjects}}{0.455} \tag{2.4}$$

In the presence of LD, we can adjust T_1 using the subject inflation factor λ_1 by

$$\frac{1}{\lambda_1} T_1 = \frac{(S_1 - E(S_1))^2}{\lambda_1 \sum_{i=1}^{m} Var(\Delta X_i)} \sim \chi^2_{(1)} \tag{2.5}$$

The adjusted test statistic T_2 is derived in the same way. Under the null-hypothesis that the subject is from the study population, the test statistics T_1 and T_2 have an asymptotic χ^2-distribution with one degree of freedom.

2.2.3 The optimal test and its asymptotic distribution

Since, prior to the calculation of the test statistic, we do not have any knowledge whether test statistic T_1 or T_2 is more suitable for the analyzed study subject, we define the genome-wide test statistic to detect genetic outliers in rare-variant data as:

$$T_{opt} = max(T_1, T_2) \tag{2.6}$$

We already know that assuming no LD between the markers, and under the null hypothesis that the subject under study is from the given population, the standardized test statistics T_1 and T_2 follow a $\chi^2_{(1)}$ distribution asymptotically. To derive the asymptotic distribution of T_{opt}, we need to incorporate the correlation between the test statistics T_1 and T_2. In the absence of LD between the genetic loci, an estimator of the correlation

between R_1 and R_2 based on the allele frequencies of the loci can be easily derived (Appendix A.2). As an alternative approach or in the presence of LD, the correlation between R_1 and R_2 can also be estimated by the empirical correlation between the statistics R_1 and R_2 in the study (Appendix A.2). Given the estimate for the correlation/covariance of R_1 and R_2, the asymptotic distribution of T_{opt} can be obtained under the null hypothesis, by simulating from a bivariate normal distribution with the estimated correlation. In Appendix A.2, we outline the derivation of the asymptotic distribution for T_{opt} in more details.

2.3 Results

We examined the performance of the test statistic T_{opt} by its applications to the third release of HapMap 3 data and the third version of 1000 Genome Project data, and in simulation studies with sequencing and GWAS data. In all applications and simulation scenarios, the approach was compared to the outlier detection algorithm based on PCA. For this comparison, we selected the smartpca implementation of PCA in the EIGENSTRAT package [Price et al., 2006].

2.3.1 Applications to HapMap 3 data

HapMap 3 [Altshuler et al., 2010] provides a unique framework to validate our approach based on real data. The genotype data were generated from 1,397 samples in 11 populations, obtained with the Affymetrix Human SNP array 6.0 and the Illumina Human1M-single Beadchip. The consensus number of polymorphic SNPs in the 11 populations is 1,457,897. We selected the US Utah residents with ancestry from northern and western Europe (CEU), Han Chinese in Beijing (CHB), and the Yoruba in Ibadan, Nigeria (YRI) as the populations for the construction of our "toy"-data sets. Since these three populations can be considered to be genetically homogeneous (Supplementary information of [Altshuler et al., 2010]), they are an ideal validation tool for methodology

to detect population substructure. The general idea is to create data sets that consist of one population, and include one additional subject that is not part of the population. The following combinations/data sets can thereby be constructed:

1. 1 CEU subject + all YRI subjects (112 such datasets)
2. 1 YRI subject + all CEU subjects (147 such datasets)
3. 1 CHB subject + all YRI subjects (137 such datasets)
4. 1 YRI subject + all CHB subjects (147 such datasets)
5. 1 CEU subject + all CHB subjects (112 such datasets)
6. 1 CHB subject + all CEU subjects (137 such datasets)

Both methods, the proposed test statistic and the outlier detection method based on PCA, were applied to each dataset. For each subset, several QC steps were implemented to ensure that the remaining autosomal SNPs have a call rate $> 98\%$, are in Hardy-Weinberg equilibrium (HWE-test p-value > 0.000001), and the subjects are unrelated. Before applying the PCA approach, SNPs in the long-range LD regions [Price et al., 2008] were also removed and the SNP-set was pruned to have pairwise $r^2 < 0.1$ in every 200 SNPs window with a step size of 20 SNPs using PLINK [Purcell et al., 2007].

In each application/replicate, we assess whether the two methods correctly identify the subject that is not part of the population as an outlier. A subject is rejected as an outlier if its test statistic T_{opt} is greater than the value corresponding to the significance level $0.05/n$ where n is the number of subjects in the dataset. In the smartpca algorithm provided in EIGENSTRAT, we used the default values recommended in the package, i.e. 10 for the number of principal components used for determining outliers, and 6 for the number of standard deviations of which the subject must deviate in any of the top 10 PCs to be removed as an outlier. We also used the default maximum number of outlier removal iterations, which is 5 in the process.

The number of false positive findings is recorded as well. Based on these results, we estimated the power, the type I error and the family-wise error rate (FWER) of both

approaches for all 6 scenarios shown above. The type I error is the average percentage of incorrectly rejected subjects among the combined datasets for each scenario. The FWER is the percentage of times that there is at least one incorrectly rejected subjects in the 112 datasets. The methods were applied first to all the common SNPs, which, for PCA, are SNPs with minor allele frequency $> 10\%$ and which, for our approach, are all the available SNPs, including rare SNPs and SNPs in the long-range LD-regions ([Price et al., 2008]). Then, we applied the two approaches to the rare SNPs (minor allele frequency $<$ 5%). The results are shown in Table 2.1.

Table 2.1: The estimated Family-wise error rate (FWER), the average type I error (TI) and the power of T_{opt} and the outlier detection process based on PCA when they were applied to the combined HapMap 3 datasets

Estimates	Pop Outlier	CEU YRI	YRI CEU	YRI CHB	CHB YRI	CEU CHB	CHB CEU
PCA (MAF $> 10\%$)	FWER	1.00	1.00	1.00	0.00	1.00	0.00
	TI	0.0178	0.107	0.106	0.00	0.0177	0.00
	POWER	1.00	1.00	1.00	1.00	1.00	1.00
PCA (MAF $< 5\%$)	FWER	1.00	1.00	1.00	1.00	1.00	1.00
	TI	0.0739	0.161	0.0752	0.158	0.0734	0.0736
	POWER	1.00	1.00	1.00	1.00	1.00	1.00
T_{opt} (MAF $< 5\%$)	FWER	0.00	0.00	0.00	0.00	0.255	0.00
	TI	0.00	0.00	0.00	0.00	0.00226	0.00
	POWER	1.00	1.00	1.00	1.00	1.00	1.00
T_{opt} (all SNPs)	FWER	0.00	0.00	0.00	0.00	1.00	0.00
	TI	0.00	0.00	0.00	0.00	0.00885	0.00
	POWER	1.00	1.00	1.00	1.00	1.00	1.00

For both common and rare variants, the results show that the outlier detection algorithm implemented in EIGENSTRAT and our approach, T_{opt}, are able to detect all the outliers, i.e. empirical power estimates of 100%, in all of the combined datasets. However, PCA is not able to maintain the type I error and the family-wise error rate (FWER). For example, the average type I error of PCA can reach levels of up to 0.1 in two of the six cases and the FWER is always 1. Our method maintains the type I error in almost all the cases, and for both rare and common variants. The exception may be due to one labelled CEU subject that is genetically far from the CEU population, so that it is often seen as an outlier of

the CEU population. Since the exclusion of study subjects with false-positive test results will reduce the statistical power of the sequence analysis or the GWAS, a large false-positive rate is not a desirable feature of a method for the detection of population outliers.

It may be argued that the criteria of detecting outliers we used in PCA may not be optimal to reject the outliers in these cases. However, we note that 6 standard deviations is commonly used in detecting outliers in the QC step and it is already a stringent criteria. The point is that the outlier detection algorithm based on PCA is not a statistical test so that it is likely to reject subjects incorrectly.

2.3.2 Applications to 1000 Genome Project data

Similarly, we applied the novel test to the 3rd release of the variant call set based on both low coverage and exome whole genome sequence data from the 1000 Genome Project [The 1000 Genome Project Consortium, 2010]. The release contains the genotype calls of 1,092 samples from 14 different populations. We combined three pairs of populations as for the HapMap 3 data to investigate the power, type I error, and FWER of the test. The three pairs are Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT), Tuscany in Italy (TSI) and Finnish from Finland (FIN), Yoruba in Ibadan, Nigeria (YRI) and Luhya in Webuye, Kenya (LWK). Each pair of the populations are genetically closer comparing to the pair of populations studied in the HapMap 3 data.

We focused only on the SNPs calls, thus any information on the short Indels or large deletions are ignored. With similar quality control process as for the HapMap 3 data, we are left with approximately 11M SNVs for the combined datasets CHB and JPT, and FIN and TSI, and with approximately 19M SNVs for the combined datasets LWK and YRI. To apply PCA, the three combined datasets, CHB and JPT, FIN and TSI, and LWK and YRI, have been pruned to include SNPs with $MAF > 10\%$ and with pairwise $r^2 < 0.05$ in each 50 SNPs window with a step size of 5 SNPs. This pruned dataset for CHB and JPT includes about 92K SNPs, similar for FIN and TSI. The pruned dataset for LWK and

YRI includes about 150K SNPs. To compare with the new test, PCA was also applied to the SNVs with MAF $\leq 5\%$ without any LD pruning. There are about 6M-7M such SNVs for the combined datasets of CHB and JPT, and of FIN and TSI, and there are about 13M such SNVs (MAF $\leq 5\%$) for the combined datasets of LWK and YRI.

The power, type I error and FWER estimates are shown in Table 2.2. In each case, there is only one outlier included in each dataset here. The power, type I error and FWER are averaged across the datasets with different outliers from the other population. The table shows that PCA cannot detect the outlier using the pruned SNP set with MAF $> 10\%$ due to the small number of SNPs included in the pruned data and the closeness of the two populations. PCA has a good power to detect the outlier using SNPs with MAF $\leq 5\%$. However, the outlier detection algorithm based on PCA does not control for the type I error or the FWER, which would result in the unnecessary removal of samples. The new statistic T_{opt} has a good power to detect the outliers, especially for the more distant pairs, TSI and FIN, and LWK and YRI. The type I error and the FWER are mostly controlled well. Note that there are a few surprises here. one is that the asymmetry in the power for the dataset of LWK with one YRI sample as the outlier and the dataset of YRI with one LWK as the outlier. This can be explained by the larger genetic variation of the LWK population than the YRI population. However, we observe that, using SNPs with $MAF \leq 5\%$, we have a much better power to detect the YRI outliers included in the LWK samples. This may due to the fact that a lot of variants that contribute in distinguishing the two populations are rare since the separation of the two populations are relatively recent. This is also true for the other combined datasets.

We further investigated the performance of the test by introducing more than one outliers into the dataset. The results are shown in Table 2.3 and Table 2.4. We randomly selected 5 or 10 outliers from the outlier population and they were combined with the corresponding study population to assess the performance of the approaches. There are 1000 such randomly generated datasets in all the scenarios except for evaluating the performance of PCA on SNVs with MAF $\leq 5\%$, where 500 datasets were generated.

Table 2.2: The estimated Family-wise error rate (FWER), the average type I error (TI) and the power of T_{opt} and the outlier detection process based on PCA when they were applied to the combined 1000 genome datasets

	Pop	CHB	JPT	TSI	FIN	LWK	YRI
Estimates	Outlier	JPT	CHB	FIN	TSI	YRI	LWK
PCA (MAF $>$ 10%)	FWER	0.00	0.00	0.00	0.00	1.00	0.00
	TI	0.00	0.00	0.00	0.00	0.0125	0.00
	POWER	0.00	0.00	0.151	0.00	0.0349	0.00
PCA (MAF $<$ 5%)	FWER	1.00	1.00	0.151	0.990	1.00	1.00
	TI	0.144	0.0935	0.00351	0.0765	0.0489	0.0234
	POWER	0.843	0.443	0.957	0.888	0.570	1.00
T_{opt} (MAF $<$ 5%)	FWER	0.00	1.00	0.00	0.00	0.00	0.00
	TI	0.00	0.0415	0.00	0.00	0.00	0.00
	POWER	0.146	0.495	1.00	1.00	0.988	1.00
T_{opt} (all SNPs)	FWER	0.0225	1.00	0.882	0.00	0.00	0.00
	TI	0.00	0.0365	0.000869	0.00	0.00	0.00
	POWER	0.0562	0.0928	0.720	0.969	0.00	0.861

We observe that even with more outliers included in the dataset, our method performs generally better than PCA, especially in the combined datasets of TSI and FIN, and LWK and YRI. PCA continues to have a large type I error rate and FWER in all the scenarios. We also observe again that the performance of the novel test is much better using the SNPs with MAF $\leq 5\%$, than using all the SNPs.

Note that as the proportion of outliers included in the dataset continues to increase, the power of our test decreases. This is because that as more outliers are included in the dataset, the estimated MAF and the expected number of alleles obtained from the data are biased toward the outlier population. Then the test statistics would be biased and would not follow the same distribution as under the null hypothesis. Thus, the novel method is mainly used to detect a relatively small set of outliers. The effect of the proportion of the outliers included in the dataset on the test statistics also depends on the genetic distance of the populations.

Table 2.3: The estimated Family-wise error rate (FWER), the average type I error (TI) and the power of T_{opt} and the outlier detection process based on PCA when they were applied to the combined 1000 genome datasets with 5 outliers included in each scenario

	Pop	CHB	JPT	TSI	FIN	LWK	YRI
Estimates	Outlier	JPT	CHB	FIN	TSI	YRI	LWK
PCA (MAF $>$ 10%)	FWER	0.00	0.0540	0.00	0.00	1.00	0.00
	TI	0.00	0.000607	0.00	0.00	0.0276	0.00
	POWER	0.01	0.00	0.00	0.00	0.00	0.00
PCA (MAF $<$ 5%)	FWER	1.00	0.984	0.900	0.820	0.982	0.926
	TI	0.125	0.0939	0.0509	0.0271	0.0531	0.0219
	POWER	0.685	0.248	0.166	0.391	0.002	0.919
T_{opt} (MAF $<$ 5%)	FWER	0.00	1.00	0.00	0.00	0.00	0.00
	TI	0.00	0.0119	0.00	0.00	0.00	0.00
	POWER	0.0728	0.301	1.00	1.00	0.988	1.00
T_{opt} (all SNPs)	FWER	0.001	1.00	1.00	0.00	0.00	0.00
	TI	0.0000103	0.0251	0.0102	0.00	0.00	0.00
	POWER	0.0552	0.0628	0.727	0.480	0.00	0.485

Table 2.4: The estimated Family-wise error rate (FWER), the average type I error (TI) and the power of T_{opt} and the outlier detection process based on PCA when they were applied to the combined 1000 genome datasets with 10 outliers included in each scenario

	Pop	CHB	JPT	TSI	FIN	LWK	YRI
Estimates	Outlier	JPT	CHB	FIN	TSI	YRI	LWK
PCA (MAF $>$ 10%)	FWER	0.00	0.166	0.00	0.00	0.746	0.00
	TI	0.00	0.0168	0.00	0.00	0.0158	0.00
	POWER	0.00	0.00	0.00	0.00	0.00	0.0024
PCA (MAF $<$ 5%)	FWER	1.00	0.810	0.966	0.676	0.978	0.985
	TI	0.100	0.0723	0.0230	0.0151	0.0500	0.0352
	POWER	0.479	0.127	0.854	0.258	0.0618	0.885
T_{opt} (MAF $<$ 5%)	FWER	0.00	0.0119	0.00	0.00	0.00	0.00
	TI	0.00	0.000134	0.00	0.00	0.00	0.00
	POWER	0.0286	0.204	0.982	0.0846	0.633	1.00
T_{opt} (all SNPs)	FWER	0.00	1.00	0.982	0.00	0.00	0.00
	TI	0.00	0.0207	0.0100	0.00	0.00	0.00
	POWER	0.0410	0.0421	0.459	0.0557	0.00	0.124

2.3.3 Simulations

In addition to the analyses on the HapMap 3 and 1000 Genome Project data, we performed simulation studies under the alternative hypothesis to examine whether the proposed test T_{opt} has sufficient power to detect outliers. We assessed the power of the approach to detect genetic outliers based on both rare variants and common variants. In our simulations, the Balding-Nichols model [Balding and Nichols, 1995] was applied to generate the allele frequencies of the two sub-populations: $p_{i1}, p_{i2} \sim Beta(\frac{1-F}{F}p_i, \frac{1-F}{F}(1-p_i))$, where p_{i1}, p_{i2} are the allele frequencies of marker locus i for the two sub-populations; F is the F_{st}, the genetic distance between the two sub-populations [Holsinger and Weir, 2009] and the parameter p_i is the background allele frequency for marker locus i. In the simulation studies for the rare variants, the background allele frequencies p_i were generated from the Wright's distribution [Wright, 1949] using Metropolis-Hastings algorithm: $f(p) = cp^{\beta_s - 1}(1-p)^{\beta_n - 1}e^{\sigma(1-p)}$, where the scaled mutation rates are elected to be $\beta_s = 0.001, \beta_n = \beta_s/3$, the selection rate $\sigma = 12$, and c is a normalizing constant. The Wright's distribution is expected to simulate the MAF spectra of the human genome under week purifying selection, where most of the MAFs generated are smaller than 5% [Wright, 1949]. To generate common variant data, we generated the background allele frequencies p_i from the Uniform distribution $Unif(0, 0.5)$.

Under the models defined by these parameters, we draw two sets of allele frequencies for the two sub-populations in each trial. In analogy to the HapMap analysis, one study subject was generated from the first sub-population, while the remaining study subjects in the dataset were generated from the second sub-population.

Power

Based on 1000 replicates, we estimated the power of the test under each scenario for both common variant and rare variant data. The power is estimated by the percentage of trials in that the outlier is detected using T_{opt}, where the test statistic is adjusted for multiple comparisons, i.e. study subjects, using the Bonferroni correction. The results

are shown in Table 2.5 and Table 2.6. Table 2.5 suggest that the test T_{opt} has sufficient power in most of the scenarios for rare variant data, especially for datasets with one million markers. This is expected since, as the number of markers increases, there is more information about the genetic structure of the population, and it is easier for the test to capture any small difference between the outlier and the rest of the subjects in the dataset. The genetic distance of the two sub-populations, F_{st}, is varied over a wide range, and we observe a decrease in the power of the test as F_{st} decreases, as we would expect. Also, we find that the percentage of markers with smaller MAF in the first population than in the second population also influences the power of the test, especially when the two populations are genetically close to each other. However, after assessing the power for the percentage between 50% and 75% (data not shown), we found that, as long as the percentage is above 50%, i.e. there is LD in the sample, we have a good power to detect the outlier under the scenarios we considered.

For common variant data, the power of the test becomes very small when the percentage of markers with smaller MAF in one population than in the other population is approximately 50%, even for large number of SNPs. However, as the percentage increases. the power increases rapidly. This is due to the small power offered by the score S_2 since S_1 does not have much power under the 50% scenario. It is important to note that, in a real data set, we would not expect the percentage to be exactly 50% if all available genetic loci are included in the calculation of the test statistic and there is LD between the loci.

We also compared our approach and the PCA approach for rare variant data, as shown in Table 2.7. We observe that both the proposed test statistic and PCA have sufficient statistical power in most scenarios. However, when there is a systematic difference in allele frequencies between the two populations, the PCA approach does not perform well. In practice, this effect on PCA can be minimized by the removal of long-range LD-regions and LD-pruning for common variant analysis, but would be unavoidable for sequence data.

Table 2.5: Power of T_{opt} for rare variant data

F_{st}	Perc	500*10k	1000*10k	500*100k	1000*100k	500*1M	1000*1M
0.20	100%	0.898	0.905	0.991	0.990	0.998	1.000
	75%	0.879	0.867	0.990	0.995	0.997	0.998
	50%	0.914	0.870	0.988	0.986	1.000	1.000
0.15	100%	0.904	0.900	0.983	0.987	1.000	0.998
	75%	0.854	0.857	0.989	0.954	0.999	0.998
	50%	0.880	0.856	0.987	0.987	1.000	1.000
0.10	100%	0.894	0.887	0.988	0.984	1.000	1.000
	75%	0.859	0.833	0.981	0.982	0.998	0.999
	50%	0.835	0.796	0.981	0.986	1.000	0.999
0.05	100%	0.878	0.875	0.980	0.983	0.997	0.996
	75%	0.807	0.777	0.973	0.974	0.998	0.997
	50%	0.388	0.354	0.967	0.970	0.998	0.996
0.01	100%	0.828	0.825	0.973	0.979	0.999	0.999
	75%	0.453	0.401	0.968	0.963	0.995	0.997
	50%	0.003	0.006	0.031	0.025	0.863	0.839
0.005	100%	0.757	0.748	0.987	0.977	0.997	0.999
	75%	0.183	0.119	0.947	0.950	0.993	0.997
	50%	0.005	0.003	0.002	0.004	0.149	0.118

The number of subjects in the datasets is either 500 or 1000. The number of SNPs included is 10,000, 100,000, or 1 million. The first column refers to the genetic distance of the two sup-populations in the dataset. The second column shows the percentage of the markers with a smaller MAF in the first sub-population than in the second sub-population.

Table 2.6: Power of T_{opt} for common variant data

F_{st}	Perc	500*10k	1000*10k	500*100k	1000*100k	500*1M	1000*1M
0.20	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	1.000	1.000	1.000	1.000	1.000	1.000
	50%	0.003	0.004	0.001	0.002	0.003	0.006
0.15	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	1.000	1.000	1.000	1.000	1.000	1.000
	50%	0.001	0.002	0.002	0.001	0.001	0.001
0.10	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	1.000	1.000	1.000	1.000	1.000	1.000
	50%	0.001	0.000	0.003	0.001	0.002	0.001
0.05	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	1.000	1.000	1.000	1.000	1.000	1.000
	50%	0.003	0.000	0.000	0.000	0.000	0.000
0.01	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	0.999	0.999	1.000	1.000	1.000	1.000
	50%	0.056	0.030	0.000	0.000	0.000	0.000
0.005	100%	1.000	1.000	1.000	1.000	1.000	1.000
	75%	0.901	0.880	1.000	1.000	1.000	1.000
	50%	0.009	0.011	0.034	0.017	0.001	0.000

The ancestral MAFs are generated from Unif(0, 0.5) for the datasets. The number of subjects in the datasets is either 500 or 1000. The number of SNPs included is 10,000, 100,000, or 1 million. The first column refers to the genetic distance of the two sup-populations in the dataset. The second column shows the percentage of the markers with a smaller MAF in the first sub-population than in the second sub-population.

Table 2.7: Power of T_{opt} and the outlier detection process based on PCA for rare variant data.

F_{st}	Perc	500 subjects x 10 k		1000 subjects x 10 k	
		PCA	T_{opt}	PCA	T_{opt}
0.2	100%	0.05	0.93	0.00	0.90
	75%	0.90	0.84	0.90	0.87
	50%	0.93	0.87	0.93	0.87
0.15	100%	0.06	0.92	0.05	0.87
	75%	0.86	0.88	0.90	0.79
	50%	0.94	0.90	0.96	0.92
0.10	100%	0.03	0.96	0.05	0.88
	75%	0.65	0.84	0.67	0.81
	50%	0.92	0.80	0.94	0.86
0.05	100%	0.03	0.89	0.02	0.88
	75%	0.20	0.80	0.38	0.77
	50%	0.90	0.33	0.94	0.28
0.01	100%	0.04	0.84	0.04	0.78
	75%	0.05	0.37	0.04	0.37
	50%	0.37	0.00	0.48	0.00
0.005	100%	0.00	0.73	0.00	0.73
	75%	0.03	0.15	0.07	0.12
	50%	0.10	0.00	0.19	0.01

The ancestral MAFs are generated from the Wrights distribution for the datasets. The number of SNPs included is 10,000.

Type I Error

Using the same set of simulated data, the type I error is estimated as the average percentage of subjects who are incorrectly rejected. The results for the rare variants are shown in Table 2.8. In the scenarios considered, the nominal type I error is $0.05/n$, where n is the number of subjects included in the dataset, to maintain the FWER at 0.05 level. Thus the nominal type I error is 0.0001 for 500 subjects and 0.00005 for 1000 subjects. From the results, we observe that for rare SNPs, the type I error for 10000 SNPs is inflated, but for datasets with a large number of SNPs, the type I error rate is acceptable. For common variants, the type I error is well-maintained in all the scenarios (data not shown).

Table 2.8: Type I error of T_{opt} for rare variant data

F_{st}	Perc	500*10k	1000*10k	500*100k	1000*100k	500*1M	1000*1M
0.20	100%	0.00190	0.00235	0.000611	0.000267	0.0000842	0.0000390
	75%	0.00247	0.00206	0.000291	0.000167	0.0000922	0.0000551
	50%	0.00161	0.00255	0.000329	0.000410	0.0000782	0.0000240
0.15	100%	0.00189	0.00155	0.000517	0.000625	0.0000581	0.0000270
	75%	0.00229	0.00257	0.000679	0.000381	0.000108	0.0000561
	50%	0.00273	0.00241	0.000361	0.000128	0.0000802	0.0000300
0.10	100%	0.00211	0.00265	0.000475	0.000332	0.0000782	0.0000350
	75%	0.00221	0.00239	0.000293	0.000150	0.000164	0.0000440
	50%	0.00362	0.00211	0.000792	0.000289	0.000190	0.0000430
0.05	100%	0.00274	0.00190	0.000569	0.000261	0.0000902	0.000181
	75%	0.00337	0.00296	0.000754	0.000290	0.000118	0.000182
	50%	0.00287	0.00288	0.000443	0.000432	0.000132	0.0000741
0.01	100%	0.00326	0.00228	0.000523	0.000420	0.000122	0.0000821
	75%	0.00296	0.00272	0.000361	0.000673	0.000152	0.0000881
	50%	0.00370	0.00359	0.000627	0.000186	0.000140	0.0000671
0.005	100%	0.00226	0.00211	0.000387	0.000242	0.000325	0.0000551
	75%	0.00322	0.00261	0.000291	0.000124	0.000204	0.0000801
	50%	0.00350	0.00367	0.000409	0.000713	0.000184	0.0000631

The ancestral MAFs are generated from the Wrights distribution for the datasets. The number of subjects in the datasets is either 500 or 1000. The number of SNPs included is 10,000, 100,000, or 1 million. The first column refers to the genetic distance of the two sup-populations in the dataset. The second column shows the percentage of the markers with a smaller MAF in the first sub-population than in the second sub-population.

The same pattern is observed for the FWER as for the type I error rate. FWER is estimated as the number of trials among 1000 trials such that at least one subject is wrongly rejected. For common variant data, the FWER is well below 0.05 for all the scenarios. For rare variant data, we do see an inflation in the FWER as the type I error rate when the number of SNPs included in the dataset is small. However, in real data set, e.g. whole exome sequencing, GWAS, etc., we expect that a sufficient number of loci is available to guarantee that the FWER is maintained.

As a last comparison, we assessed the performance of both approaches under the null hypothesis. As shown in Table 2.9, for rare variants generated from the Wright's distribution, PCA has much larger FWER compared to T_{opt}. In almost all the trails, PCA rejected at least one subject incorrectly, whereas T_{opt} has been shown above that the FWER is acceptable when the number of SNPs included is sufficiently large. For common variants, the FWER of both approaches is well-maintained with 500 or 1000 subjects included in the datasets.

Table 2.9: FWER of T_{opt} and the outlier detection process based on PCA for rare variant data

FWER	500 x 10 k		1000 x 10 k		500 x 100 k		1000 x 100 k	
Dist	PCA	T_{opt}	PCA	T_{opt}	PCA	T_{opt}	PCA	T_{opt}
Wright	0.912	0.106	0.978	0.123	0.928	0.066	0.989	0.060

2.4 Discussion

The large-scale applications of next-generation sequencing technology to association studies require the development of robust and powerful analysis approaches. While substantial progress has been made in terms of the development of association tests for rare variants [Li and Leal, 2008] [Madsen and Browning, 2009] [Ionita-Laza et al., 2011] [Mukhopadhyay et al., 2010] [Neale et al., 2011], there is yet no standard statistical

approach that addresses the issues of population-substructure for sequence data.

Recently, a permutation procedure is proposed by Epstein et al [Epstein et al., 2012] to address the problem in association tests of rare variations. It is a nice approach that benefits the rare-variant association tests that cannot correct for confounding. However, it is subject to the same problem as other rare-variant association tests that may be adjusted for ancestry due the fact that the ancestry covariates obtained using PCA may not be accurate as the type I error of the association tests after adjusting for ancestry using PCA has been shown to be still inflated under certain scenarios [Mathieson and McVean, 2012]. Here, we try to approach the problem from a different direction, by obtaining a homogeneous sub-population to remove confounding and avoid the hassle of estimating the ancestry covariates for rare variants. In this communication, we proposed a method that can detect study subject that introduce population substructure in the sample, potentially confounding the association analysis. Our approach is computationally fast and simple, i.e. the method is computed based on all available genetic loci, making LD estimation and pruning unnecessary. The approach works well for both rare and common variants. We illustrated this by the applications to the HapMap 3 and the 1000 Genome Project data, and in our simulation studies.

While these are the advantages over the standard PCA analysis, our approach does not assess the pairwise similarity of study subjects, e.g. principal component plots. This restricts our approach to the role of an outlier detection tool. Unlike principal components, an integration of the test statistic into a regression model as an adjustment for population substructure is problematic for this reason. Additional research on this topic is required.

On the simultaneous association analysis of large genomic regions: A massive multi-locus association test

Dandi Qiao, Heide Fier, Michael Cho, Edwin K. Silverman, Christoph Lange

Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

Department of Genomic Mathematics, University of Bonn, Bonn, Germany

3.1 Introduction

In the search for disease susceptibility loci (DSLs), genome-wide association studies (GWAS) have been a successful instrument for the identification of replicable genetic associations [Manolio et al., 2008, Hardy and Singleton, 2009]. For most complex diseases and phenotypes, they discovered numerous genetic associations that can be validated in independent populations, although the genetic effect sizes of the loci are relatively small. Despite of the large number of detected loci, GWAS association signals are only able to explain a small fraction of the overall predicted heritability [Visscher et al., 2008], i.e. the phenomenon of "missing heritability". One possible explanation for this phenomenon is "synthetic associations" [Dickson et al., 2010]. Simulation studies, theoretical considerations and empirical evidence [Nejentsev et al., 2009, Adzhubei et al., 2010] , suggest that genetic associations, as they are detected by GWAS, can be caused by multiple rare variants (RVs). Because common variants are poor proxies for RVs or are not in linkage disequilibrium with rare disease-causing variants, it is difficult to identify or characterize rare DSLs in GWAS data.

Another plausible explanation for the phenomenon of "missing heritability" is insufficient statistical power due to the multiple-testing problem. In a GWAS, one million and more genetic loci are tested individually for association with the target phenotype, and the test results have to be adjusted for multiple comparisons, leading to extremely small p-value thresholds for overall statistical significance. The standard approach has been aimed to increase the sample size of GWAS as much as possible. For example, several meta-analyses of GWASs [Allen et al., 2010] have contained the data of more than 100,000 study subjects. However, such large sample sizes hold the danger of increased study heterogeneity and do not necessarily lead to increased statistical power.

The fundamental issue with the standard analysis approach to GWAS (single locus association testing and adjustment for multiple comparisons), is that an increase in genomic resolution, i.e. adding more and more genetic loci to the analysis, does not

increase the probability to detect DSLs, but diminishes the statistical power of the approach. To address this issue, multi-loci tests have been suggested. For example, gene-based analysis has been advocated [Neale and Sham, 2004] to complement allelic association analysis of single loci. This is motivated by the idea that causal variants for one disease tend to reside in proximity to each other and variants in adjacent regulatory regions are more likely to have functional relevance [Huang et al., 2011]. PLINK [Purcell et al., 2007] provides ”set-based” tests using the average SNP statistic across the set of SNPs to realize this idea. Moreover, other tests such as the minSNP test, the Bayesian imputation-based association mapping (BIMBAM) test [Servin and Stephens, 2007], the versatile gene-based test (VEGAS) test [Liu et al., 2010] and the LASSO regression method for GWAS [Wu et al., 2009] has been proposed. Recently, the Gene-wide Significance (GWiS) test [Huang et al., 2011] has also been developed by Huang et al which could also be used to estimate the number of independent effects within a gene. For next-generation sequencing data, methods aggregate over a set of rare variants to search for associated genomic regions with the disease status are shown to be more powerful, such as the cohort allelic sums test (CAST) [Morgenthaler and Thilly, 2007], the Combined Multivariate and Collapsing (CMC) method [Li and Leal, 2008], the weighted sum statistic by Madsen and Browning [Madsen and Browning, 2009], the kernal-based adaptive clustering (KBAC) test [Liu and Leal, 2010], the sequence kernel association test (SKAT) [Wu et al., 2011], replication-based test (RBT) [Ionita-Laza et al., 2011], etc. There are several advantages of such gene-based tests over single loci tests. First, collapsing the small effects across the variants within a gene could give larger effect size to detect the association. Secondly, due to the smaller number of genes to be tested, the multiple testing problem is reduced. Moreover, the associations of genes across different populations can be directly compared even though there could be different linkage disequilibrium patterns within the genes across the populations [Huang et al., 2011].

However, all of the approaches can handle only a very limited number of genetic loci, i.e. typically less than 100. None of them is able to incorporate the information about the

physical location of the loci and their clustering. In this paper, we are proposing a novel approach that can test a large genomic region for association with the target phenotype by taking into account the physical location of the variants that show evidence for association and their physical clustering. The genomic region could refer to one gene, a specified segment of the genome, a pathway, an entire chromosome or the complete genome. The approach is computationally fast and applicable to binary and complex phenotypes. The methodology is evaluated in simulation studies and by applications to a GWAS dataset from the COPDgene study. The simulation studies suggest that the approach has sufficient power to test simultaneously all genotyped loci on the entire genome or a specific chromosome.

3.2 Methods

The proposed test assesses whether there is significant clustering of causal variants within a specified region. We consider both the level of associations between the variants and the trait, and the location of the variants. The degree of association between a variant and the phenotype is represented by the association p-values, which is easy to obtain from any dataset and allows the application of our method to both quantitative traits and dichotomous traits. To put this into a one-dimensional clustering problem, we need to consider four aspects of the test:

1) What distance measure to use: the physical distance between two variants or a newly defined distance measure.

2) Which SNVs to look at: the cut-off value for the p-values of the variants.

3) Whether to look at the distance to the nearest neighbor or the distances to the neighboring variants, and how many neighboring variants should be considered in the calculation of distances.

4) How to quantify the difference between the distribution of the observed distances and the distribution of distances under the null, i.e. what test to use.

3.2.1 Distance measure

The first three questions shown above refer to the choice of distance distribution. Considering the absolute size of the physical distances between variants and the p-values obtained from the association tests, our goal is to have a distance measure such that the distance between two variants is small if the "average" p-value of the two variants is small, and if the physical distance between the two variants is small, relative to the other variants. Thus, we consider the multiplication of the physical distance with the association information rather than the addition of the two values to avoid the situation where the "average p-value" is overwhelmed by the physical distance. To obtain the "average" degree of association of the two variants, multiplication of the two p-values is also more suitable than addition since one large p-value would dominate a much smaller p-value. We define a new distance measure D between two variants that combines the p-value with the physical distance between the variants:

$$D_{i,j} = dist_{i,j} * \sqrt{S_i S_j}$$

where the subscript i and j refer to any two variants in the region of interest. The distance measure is motivated by the fact that this distance equals the area below the geometric average of the p-values of the two variants. Note that the distance distribution used for testing and the actual test used in our method, as described below, depend only on the relative value of the new distance measure between two variants comparing to the other distances rather than the absolute value, and the physical distance between specific variants is fixed in permutations. As a consequence, any monotonic function of $(S_i * S_j)$ (Eg. the square root) used in the definition of the distance measure will provide the same test results. We use the square root here to have the absolute value of $D_{i,j}$ to lie in a

reasonable range.

3.2.2 Cut-off values

There are two parameters that can be varied in the test: a cut-off value for the p-values – P, such that only variants with p-values below P are considered in the test of clustering; and the number of neighboring variants around each variant for calculating the distances – R. We could use a p-value of 1 to include all the variants and consider the distances from one variant to all the other variants in the region, but simulations suggest that this is computationally costly and has relatively low power comparing to including only variants with p-values below a threshold. Thus, a threshold on the p-value for selecting variants is used. The nearest neighbor method is commonly used in clustering analysis, and it requires less computational cost. However, it does not give much information on the second, third, or higher level neighbors. Thus we consider both the distance to the nearest neighbor and the distances to a pre-defined R number of neighboring variants in the region.

In our analysis, this threshold of neighboring variants R and the cut-off value of p-values P are set to be the values that correspond to specified quantiles of all the variants in the region of interest. For example, we may specify the cut-off quantile for the p-values to be 0.1%, which means the top 0.1% variants with the smallest p-values are included in the analysis. If we specify the quantile threshold of neighboring variants R to be 1%, it means that the number of neighboring variants used to calculate the distances from each variant is 1% times N, where N is the total number of variants.

3.2.3 Test on the distance distribution

To test wether there is clustering of small p-values, the distribution of the distances between the variants needs to be compared to the distribution of the distances under the null hypothesis in some way. The most popular nonparametric method to compare the

empirical distribution of one sample with a specified distribution, or to compare the empirical distributions of two samples, is the Kolmogorov-Smirnov (KS) statistic, defined as:

$$D_{n_1,n_2} = \sup_x |F_{1,n}(x) - F_{2,n}(x)|$$

where $F_{1,n}(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$ is the observed cumulative distribution function of the first sample, and similarly $F_{2,n}(x)$ is the observed cumulative distribution function of the second sample. The first sample, in our case, refers to the observed distances between the variants. The second sample, refers to the distances between the variants obtained under the null hypothesis using permutations.

We also considered an alternative approach, called the Bin test statistic as described below, that extends the idea in [Kowalski et al., 2002] [Olson et al., 2005]. The Bin test is a permutation test that compares the observed proportions of distances in ten given intervals to the expected proportions of distances using the M statistic (referred to as the Bin test):

$$M = (Prop - E(Prop))^T S^{(-1)} (Prop - E(Prop))$$

The distances between the variants obtained using permutations under the null are ordered and put into 10 bins with equal size, therefore there are 10% of all the distances in each of the 10 bins. Thus, E($Prop$) is set to be a 10×1 vector of 10% in this statistic, i.e. (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1). Then, the minimum and maximum of the distances in each bin give the interval of distance of each bin. $Prop$ is then the 10×1 vector of the proportions of the observed distances in these ten intervals. $S^{(-1)}$ is the 10×10 Moore-Penrose generalized inverse of the variance covariance matrix of the proportions of distances in the ten intervals from each permutation under the null. The number of equally spaced bins could be varied, and unequally spaced bins could be used, as discussed in [White et al., 2009]. We chose 10 equally spaced bins here to simplify the problem, but further investigation is needed to evaluate the performance of the statistic with other choices.

For both the KS and the Bin tests, the null distribution of distances is obtained by permuting the case and control status among the subjects, which conserves the linkage disequilibrium (LD) between the variants.

Other distribution tests could also be used here, such as the Ansari-Bradley test. From a limited number of simulations, the Ansari-Bradley test gives a moderate power that is higher than the KS test, but does not perform as good as the Bin test (data not shown here).

3.3 Results

We assessed the performance of the KS test and the Bin test using simulations based on the genotypes of the African American (AA) samples in the GWAS dataset of the COPDgene study [Regan et al., 2011]. Also, the Bin test was applied to the chronic obstructive pulmonary disease (COPD) status of both the African American samples and the Non-Hispanic White (NHW) samples separately. There are 682945 SNPs included for the 2569 AA samples (820 cases and 1749 controls), and 629027 SNPs included for the 5351 NHW samples (2819 cases and 2532 controls), after the quality control (QC) steps. We excluded variants with MAF < 0.01, high missing rate (above 5% for SNPs with MAF $\geq 5\%$, and above 2% for SNVs with MAF $\leq 5\%$), HWE p-value $< 10E-3$, and concordance rate $< 99\%$ using 205 duplicated samples. Samples with call rate $< 98.5\%$, and mismatched gender and race were also excluded. Autosomal SNPs with HWE p-value > 0.01, $MAF > 0.05$, and markers represented in Hapmap III were used for Principal Component analysis. Eigenstrat was used to adjust for population substructure for both AA and NHW samples to obtain the p-values of the Armitage trend test.

3.3.1 Simulation results

Simulation results on entire chromosome

In our simulations, we used the genetic data on chromosome 7 from the COPDgene study, but generated the case and control status according to our disease model. Principal components were also included to adjust for the association p-values of the SNPs in the simulation. Two different scenarios were considered. First, we selected nine SNPs on chromosome 7 as the causal variants that reside close to each other, and considered both protective and deleterious effects of the variants and different effect size. Two sets of effect sizes are simulated for this scenario. For effect 1, the odds ratio of the nine SNPs are (0.8, 1.1, 0.9, 1.2, 0.9, 1.2, 1.2, 1.5, 1.5); for effect 2, the odds ratio of the nine SNPs are (0.8, 1.1, 0.8, 1.3, 0.9, 1.3, 1.2, 1.5, 1.5). Then given the effect size and the genotypes of the samples in the COPDgene study, we generate the case and control status accordingly. Second, we considered the possibility of having a lot of causal variants with small effect size. Thus, 100 causal variants are chosen in proximity to each other on the chromosome by randomly selecting 100 variants in a randomly selected region on the chromosome. The effect sizes (odds ratio) are generated using a normal distribution with mean 1 and standard deviation 0.05.

Sensitivity analysis was done to assess the effects of the p-value cut-off P, and the number of neighbouring variants R, on the power and type I error of the Bin test. The results and discussions are in Appendix B. According to the analysis, to achieve a good power, SNPs with p-value in the top 0.5% percentile were included in the analysis, and $0.1\% * N$ neighbouring SNPs next to each SNP were used in the tests, where N is the total number of SNPs in the dataset. Due to computational limitation, 2000 permutations were used in each permutation set to maintain the type I error, as explained in our sensitivity analysis (Appendix B). For each scenario, 200 simulations were generated to obtain the estimated power and the type I error rate. The power of the test is the percentage of simulations in which the permutation p-value is less than 0.05. The results of the Bin test are shown in Table 3.1, as well as the power of the KS test, as a comparison. We observed a higher

power of the Bin test comparing to the KS test in all the scenarios. Therefore, the Bin test is recommended and is used in the calculation of the association p-values of the chromosomes in the application section.

Table 3.1: The power of the tests for three scenarios, obtained from 200 simulations with 2000 permutations in each permutation set.

	Effect1	Effect2	Effect3
Bin Test	0.920	0.990	0.345
KS Test	0.620	0.845	0.195

The power is the number of simulations with p-value less than 0.05. The effect sizes (odds ratio) of the nine SNPs with the intercept at the front are Effect 1: (0.135, 0.8, 1.1, 0.9, 1.2, 0.9, 1.2, 1.2, 1.5, 1.5) and Effect 2: (0.135, 0.8, 1.1, 0.8, 1.3, 0.9, 1.3, 1.2, 1.5, 1.5) for the first two columns. The MAF of the nine SNPs are (0.1740, 0.4914, 0.1734, 0.1244, 0.4673, 0.2552, 0.1098, 0.0309, 0.0728). For effect 3, 100 SNPs were chosen within a random segment on the chromosome and are assigned with randomly generated effect sizes from a normal distribution with mean 1 and standard deviation 0.05 with an intercept odds of 1 in each simulation.

We also computed the type I error rate of the Bin test on three different autosomal chromosomes by randomly generating the probability of having the disease for each individual using an uniform distribution $Unif(0, 0.5)$, and then randomly generated the disease status for each sample using a Bernoulli distribution with these probabilities. It is shown in Table 3.2 that the type I error rate is well-maintained with different LD patterns on different chromosomes.

Table 3.2: The type I error rate of the test on chromosome 7, 10, and 22. 2000 permutations were used and 200 replicates were generated to compute the type I error rate.

	Chromosome 7	Chromosome 10	Chromosome 22
Bin Test	0.030	0.065	0.020
KS Test	0.035	0.025	0.040

3.3.2 Application results

Results on each chromosome

The test was applied to the two subpopulations (AA and NHW) separately to see if there is any chromosome that is significantly clustered with variants associated with COPD status. Similar to the simulations, the cut-off value for the association p-value percentiles was 0.5% , and the quantile of neighbouring SNPs around each SNP to be included in the analysis was 0.1%. The p-values for testing for clustering on each chromosome were obtained using a 2000 permutation set.

Association p-values of the Armitage trend test for the SNPs, adjusted for ancestry, were computed for the AA and the NHW samples and are plotted in Figure 3.1 and 3.2 below. It has been found from previous studies that loci in the FAM13A gene on chromosome 4 is susceptible to COPD [Cho et al., 2010] and there are studies indicating that loci at the CHRNA3-CHRNA5-IREB2 locus on chromosome 15 and loci near HHIP on chromosome 4 may be related to COPD [Pillai et al., 2009] [Wilk et al., 2009]. For the AA samples, no SNP is shown to be significantly associated with COPD, as shown in Figure 3.1. However, from Figure 3.2 of the NHW samples, we observe that several SNPs reach the significance level (5×10^{-8}) on chromosome 15, but none on chromosome 4.

Thus we applied our clustering method to both datasets to see if there is any chromosome on which there is significant clustering of causal variants. The results are shown in Table 3.3. We found that for the NHW samples, both chromosome 4 and 15 are significantly clustered with causal variants, with p-values less than $0.05/22 = 0.00227$. However, in the AA dataset, we do not observe any p-value < 0.00227 for chromosome 4, which may be explained by the smaller sample size in the AA dataset.

Note that no covariate was available to us in our initial analysis. Later with information on smoking status, gender and age at enrolment included in the analysis, several SNPs on chromosome 4 and 15 are shown to be significantly associated with the COPD affection

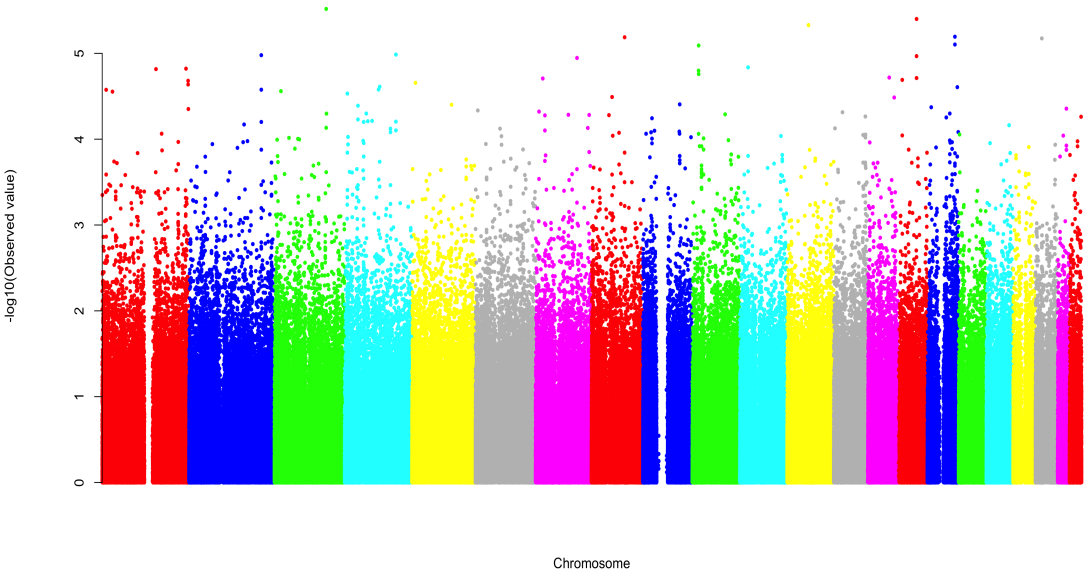

Figure 3.1: The Manhattan plot of the adjusted p-values of the SNPs in the AA dataset.

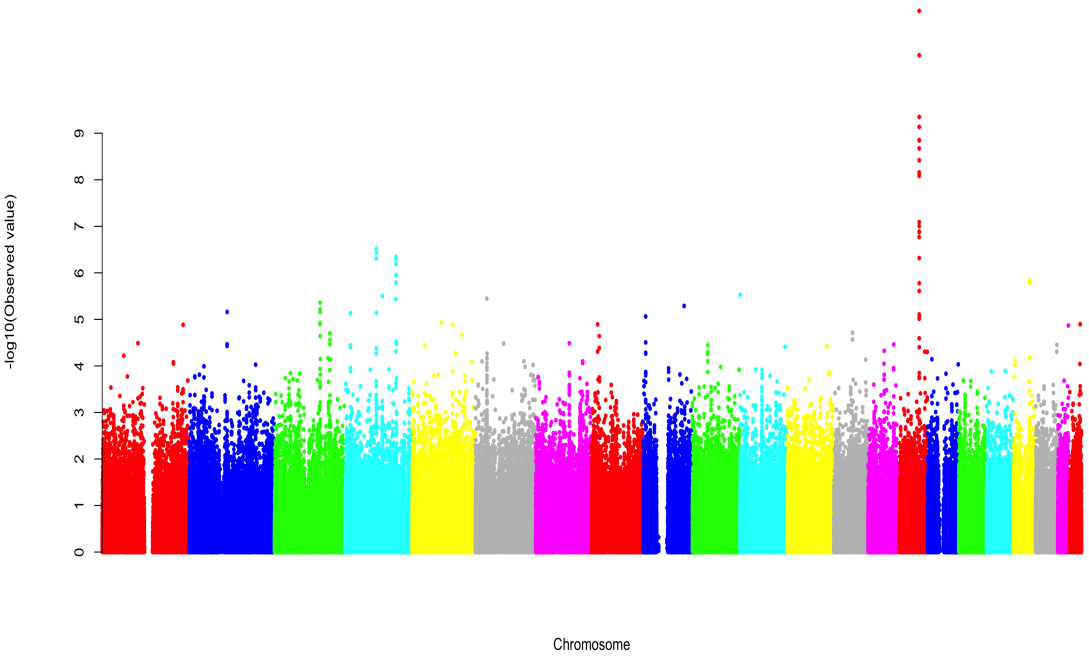

Figure 3.2: The Manhattan plot of the adjusted p-values of the SNPs in the NHW dataset.

status. However, the above analysis shows the additional significance region obtained using our method when the other information is not available.

Table 3.3: The p-value of the 22 chromosomes of the two populations in the COPDgene GWAS dataset. With Bonferroni correction, the p-values should be compared with 0.00227.

Chromosome P-value	AA (2000 perm set)	NHW (2000 perm set)
1	0.76500	0.13450
2	0.82025	0.09050
3	0.32900	0.04050
4	0.12700	0.00125
5	0.16286	0.16650
6	0.08538	0.71695
7	0.47025	0.46038
8	0.44808	0.08975
9	0.24500	0.70183
10	0.87500	0.46525
11	0.32186	0.01950
12	0.97142	0.15056
13	0.12088	0.93250
14	0.79725	0.49900
15	0.51375	0.00150
16	0.07438	0.13082
17	0.88933	0.06700
18	0.37150	0.72600
19	0.42408	0.73575
20	0.88800	0.62363
21	0.96375	0.97567
22	0.08363	0.24056

Results for genes on chromosome 4

In this section, we applied the proposed test as a gene-based test to a number of selected genes on chromosome 4. The distance distribution of chromosome 4 is shown in Figure 3.3 to compare to the distance distribution obtained using 250 permutations under the null. We observe the largest difference between the distributions at distance D around 2600. Figure 3.4 shows the physical position of

the SNPs with their p-values on the y-axis for the observed phenotype and a generated

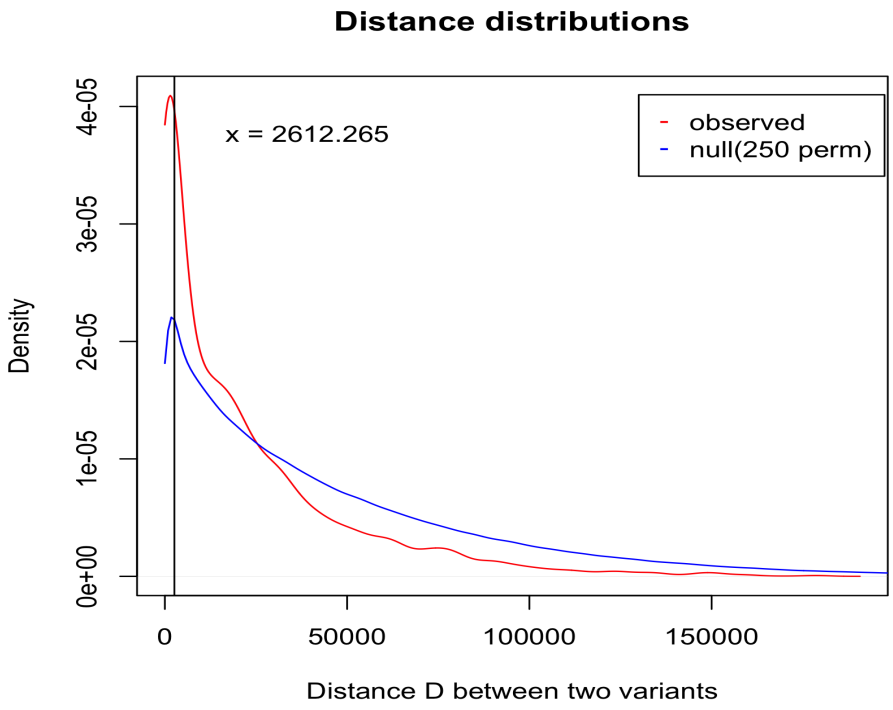

Figure 3.3: The distance distributions of the observed distances D between two variants and the distance distribution of the distances obtained using 250 permutations under the null.

phenotype from one permutation. The SNPs are colored according to the number of distances D to the neighboring SNPs that are less than 2600. The more number of neighboring SNPs with D < 2600, the deeper the red color. The two clusters shown here reside in *FAM13A* and the upstream region *HHIP* on chromosome 4q31. A less obvious cluster reside at the loci 4p15.31. To demonstrate the performance, we applied our method again on these regions and another COPD-candidate gene *PPARG* on chromosome 3. All SNPs on each gene are included and the distances to all SNPs on the gene are used. 10000 permutations are done for each case. The p-values obtained for five genes are shown in Table 3.4 and the associations of *FAM13A* and the region at 4q31.21 with COPD are shown to be significant by our method. We also test the variants in *HHIP* and found no significant association. These findings are consistent with GWAS results in lung function in COPD identifying the region upstream of *HHIP*, as well as data demonstrating a functional impact of these variants in COPD [Zhou et al., 2012].

Table 3.4: The p-value of the genes FAM13A, KRT18P51(psudogene), HHIP, PPARG and LOC729006.

Gene	FAM13A	KRT18P51	HHIP	PPARG	LOC729006
P-value	< 0.00001	< 0.00001	0.34745	0.96516	0.001

Result on the entire genome

We have also applied the test to the entire genome to see if there is any region in the genome that is clustered with the causal variants. Due to the computational cost, a much smaller p-value cut-off is used (0.025%) and a much smaller threshold for the neighboring SNP is used (0.005%). With 2000 permutations set, the p-value is < 0.0005 for the entire genome, showing strong significance of association between the genome and the phenotype. Currently, there is no other multi-loci association method that could test the association of the entire genome with the phenotype. This application shows the potential of our method for testing large genomic regions when no significant association is found for univariate tests. One possible way to search for the associated loci with the phenotype is to conduct a binary search using our method. Interested readers could refer

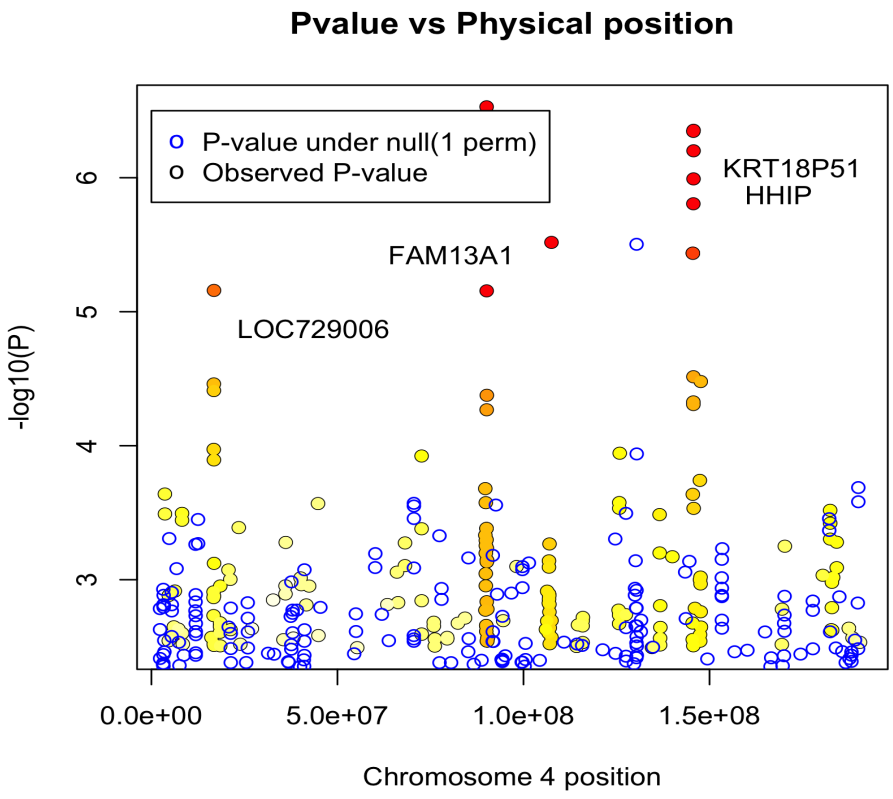

Figure 3.4: The p-values of the SNPs versus their physical positions on chromosome 4. The p-values of the SNPs from one permutation is also shown in blue circle for comparison. The black circles are colored from yellow to red according to the number of distances D that are less than 2600 between each SNP to their neighboring SNPs. The deeper the red color, the larger number of distances D that are less than 2600.

to the Appendix for some discussion about the procedure.

3.4 Discussion

In summary, we proposed here an approach for the detection of clustering of causal variants on a genomic region of any size. Many existing methods collapse the effects of variants across a region or a gene, however, very few of them utilizes the physical location of these variants and many would suffer lost of power when too many variants are included. Simulations and application results suggest that our approach provides sufficient power to detect associated genomic regions with complex disease.

The same idea of test of clustering could be applied to sequencing data, where thousands of variants would be available for each gene. For variants that are extremely rare, the univariate p-value may include only random noise, thus other measure of the association at each variant need to be considered. i.e. standard analysis approaches for rare variant analysis. Moreover, the genomic region could refer to the genes in the same pathway [Wang et al., 2007], thus whether there is significant clustering of small p-values in each pathway could be examined.

However, there are several drawbacks we need to consider. Since permutation is used to obtain the p-value of the test statistic, there is extensive computational cost if the test is applied to a large number of small regions, which requires more number of permutations. The power may also be compromised if the regions are extremely small, limiting the possibility of clusters and their detection, and if the number of regions to be tested are extremely large due to multiple testing problem. Right now the method is limited to population-based studies since permutation of the affections status is used to evaluate the p-values, but since the associations are represented by p-values, which could be obtained from either population-based association tests, or family-based association tests, there is potential to extend the approach to family-based association studies.

References

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A., and Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.

Allen, H., Estrada, K., Lettre, G., Berndt, S., Weedon, M., Rivadeneira, F., Willer, C., Jackson, A., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

Altshuler, D., Gibbs, R., Peltonen, L., Dermitzakis, E., Schaffner, S., Yu, F., Bonnen, P., de Bakker, P., Deloukas, P., Gabriel, S., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52.

Balding, D. and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1):3–12.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 11:773–785.

Burton, P., Clayton, D., Cardon, L., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D., McCarthy, M., Ouwehand, W., Samani, N., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., DeMeo, D. L., Hunninghake, G. M., Litonjua, A. A., Sparrow, D., Lange, C., Won, S., Murphy, J. R., Beaty, T. H., Regan, E. A., Make, B. J., Hokanson, J. E., Crapo, J. D., Kong, X., Anderson, W. H., Tal-Singer, R., Lomas, D. A., Bakke, P., Gulsvik, A., Pillai, S. G., and Silverman, E. K. (2010). Variants in fam13a are associated with chronic obstructive pulmonary disease. *Nat Genet*, 42(3):200–202.

Christley, S., Lu, Y., Li, C., and Xie, X. (2009). Human genomes as email attachements. *Bioinformatics*, 25:274–275.

Cohen, J., Kiss, R., Pertsemlidis, A., Marcel, Y., McPherson, R., and Hobbs, H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869.

Cohen, J., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G., Grundy, S., and Hobbs, H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proceedings of the National Academy of Sciences of the United States of America*, 103(6):1810.

Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S., et al. (2011). The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Dickson, S., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(1):e1000294.

Epstein, M. P., Duncan, R., Jiang, Y., Conneely, K. N., Allen, A. S., and Satten, G. A. (2012). A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *The American Journal of Human Genetics*, 91(2):215 – 223.

Fearnhead, N., Wilding, J., Winney, B., Tonks, S., Bartlett, S., Bicknell, D., Tomlinson, I., Mortensen, N., and Bodmer, W. (2004). Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proceedings of the National Academy of Sciences of the United States of America*, 101(45):15992.

Hardy, J. and Singleton, A. (2009). Genomewide association studies and human disease. *New England Journal of Medicine*, 360(17):1759–1768.

He, H., Zhang, X., Ding, L., Baye, T., Kurowski, B., and Martin, L. (2011). Effect of population stratification analysis on false-positive rates for common and rare variants. In *BMC proceedings*, volume 5, page S116. BioMed Central Ltd.

Helbig, I., Mefford, H., Sharp, A., Guipponi, M., Fichera, M., Franke, A., Muhle, H., De Kovel, C., Baker, C., Von Spiczak, S., et al. (2009). 15q13. 3 microdeletions increase risk of idiopathic generalized epilepsy. *Nature genetics*, 41(2):160–162.

Holsinger, K. and Weir, B. (2009). Genetics in geographically structured populations: defining, estimating and interpreting fst. *Nature Reviews Genetics*, 10(9):639–650.

Huang, H., Chanda, P., Alonso, A., Bader, J., and Arking, D. (2011). Gene-based tests of association. *PLoS genetics*, 7(7):e1002177.

Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietil
"ainen, O., Buizer-Voskamp, J., Strengman, E., Francks, C., Muglia, P., et al. (2011). Copy number variations of chromosome 16p13. 1 region associated with schizophrenia. *Molecular psychiatry*, 16:17–25.

Ionita-Laza, I., Buxbaum, J., Laird, N., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS genetics*, 7(2):e1001289.

Ji, W., Foo, J., O'Roak, B., Zhao, H., Larson, M., Simon, D., Newton-Cheh, C., et al. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature genetics*, 40(5):592–599.

Kowalski, J., Pagano, M., and DeGruttola, V. (2002). A nonparametric test of gene region heterogeneity associated with phenotype. *Journal of the American Statistical Association*, 97(458):398–408.

Kryukov, G., Pennacchio, L., and Sunyaev, S. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, 80(4):727–739.

Lange, C., Dawn, D., Edwin, K. S., Scott, T. W., and Nan, M. L. (2004). Pbat: Tools for family-based association studies. *Am J Hum Genet*, 74(2):367–369.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*, 6(10).

Liu, J., Mcrae, A., Nyholt, D., Medland, S., Wray, N., Brown, K., et al. (2010). A versatile gene-based test for genome-wide association studies. *American journal of human genetics*, 87(1):139.

Luca, D., Ringquist, S., Klei, L., Lee, A., Gieger, C., Wichmann, H., Schreiber, S., Krawczak, M., Lu, Y., Styche, A., et al. (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *The American Journal of Human Genetics*, 82(2):453–463.

Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

Manolio, T., Brooks, L., and Collins, F. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590.

Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, 5(10):e1000686.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

Morgenthaler, S. and Thilly, W. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56.

Mukhopadhyay, I., Feingold, E., Weeks, D., and Thalamuthu, A. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, 34(3):213–221.

Neale, B., Rivas, M., Voight, B., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S., Roeder, K., and Daly, M. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Neale, B. and Sham, P. (2004). The future of association studies: gene-based analysis and replication. *The American Journal of Human Genetics*, 75(3):353–362.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387.

Novembre, J. and Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649.

Olson, K. L., Bonetti, M., Pagano, M., and Mandl, K. D. (2005). Real time spatial cluster detection using interpoint distances among precise patient locations. *BMC Med Inform Decis Mak*, 5:19–19.

Patterson, N., Price, A., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190.

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., Feng, S., Hersh, C. P., Bakke, P., Gulsvik, A., Ruppert, A., Lodrup Carlsen, K. C., Roses, A., Anderson, W., Rennard, S. I., Lomas, D. A., Silverman, E. K., Goldstein, D. B., and ICGN Investigators (2009). A genome-wide association study in chronic obstructive pulmonary disease (copd): identification of two major susceptibility loci. *PLoS Genet*, 5(3).

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

Price, A., Weale, M., Patterson, N., Myers, S., Need, A., Shianna, K., Ge, D., Rotter, J., Torres, E., Taylor, K., et al. (2008). Long-range ld can confound genome scans in admixed populations. *American journal of human genetics*, 83(1):132.

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137.

Pritchard, J. and Cox, N. (2002). The allelic architecture of human disease genes: common disease-common variant... or not? *Human molecular genetics*, 11(20):2417.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.

Regan, E., Hokanson, J., Murphy, J., Make, B., Lynch, D., Beaty, T., Curran-Everett, D., Silverman, E., and Crapo, J. (2011). Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43.

Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P., Richter, D., Lavery, T., Kouyoumjian, R., Farhadian, S., Ward, R., et al. (2001a). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Reich, D., Goldstein, D., et al. (2001b). Detecting association in a case-control study while correcting for population stratification. *Genetic epidemiology*, 20(1):4–16.

Roy-Gagnon, M., Moreau, C., Bherer, C., St-Onge, P., Sinnett, D., Laprise, C., Vézina, H., and Labuda, D. (2011). Genomic and genealogical investigation of the french canadian founder population structure. *Human genetics*, 129(5):521–531.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885.

Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J., et al. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210):232–236.

Stranger, B., Stahl, E., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.

The 1000 Genome Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073.

Visscher, P., Hill, W., and Wray, N. (2008). Heritability in the genomics eraconcepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266.

Walsh, T., McClellan, J., McCarthy, S., Addington, A., Pierce, S., Cooper, G., Nord, A., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

Weiss, L., Shen, Y., Korn, J., Arking, D., Miller, D., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M., Green, T., et al. (2008). Association between microdeletion and microduplication at 16p11. 2 and autism. *New England Journal of Medicine*, 358(7):667.

White, L., Bonetti, M., and Pagano, M. (2009). The choice of the number of bins for the m statistic. *Computational statistics & data analysis*, 53(10):3640–3649.

Wilk, J. B., Chen, T. H., Gottlieb, D. J., Walter, R. E., Nagle, M. W., Brandler, B. J., Myers, R. H., Borecki, I. B., Silverman, E. K., Weiss, S. T., and O'Connor, G. T. (2009). A genome-wide association study of pulmonary function measures in the framingham heart study. *PLoS Genet*, 5(3).

Wright, S. (1949). Adaptation and selection. *Genetics, paleontology and evolution*, pages 365–389.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Wu, T., Chen, Y., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721.

Zhou, X., Baron, R., Hardin, M., Cho, M., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C., Mancini, J., Lu, K., et al. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates hhip. *Human molecular genetics*, 21(6):1325–1335.

Appendix A

Asymptotic distribution of T_{opt}

A.1 Derivation of the estimated expected value and variance of the statistics

Under the assumption of Hardy-Weinberg equilibrium, and assuming that there is no LD between the loci, the mean and variance of S_1 and S_2 can be derived analytically based on the estimated allele frequencies. They are given by:

$$E(S_1) = E(\sum_{i=1}^{m}(X_i - 2p_i)) = \sum_{i=1}^{m}(E(X_i) - 2p_i) = 0$$

$$Var(S_1) = Var(\sum_{i=1}^{m}(X_i - 2p_i)) = \sum_{i=1}^{m} Var(X_i - 2p_i) = \sum_{i=1}^{m}(2p_i(1-p_i))$$

$$E(S_2) = E(\sum_{i=1}^{m}(|X_i - 2p_i|)) = \sum_{i=1}^{m}(E(|X_i - 2p_i|)) = \sum_{i=1}^{m}(4p_i(1-p_i)^2)$$

$$Var(S_2) = Var(\sum_{i=1}^{m}(|X_i - 2p_i|)) = \sum_{i=1}^{m} Var(|X_i - 2p_i|) = \sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))$$

Then the standardized test statistics based on the estimated allele frequencies are:

$$T_1 = \frac{[\sum_{i=1}^{m}(X_i - 2p_i)]^2}{\sum_{i=1}^{m}(2p_i(1-p_i))} \tag{A.1}$$

$$T_2 = \frac{[\sum_{i=1}^{m}(|X_i - 2p_i| - (4p_i(1-p_i)^2))]^2}{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))} \tag{A.2}$$

A.2 Derivation of the correlation of R_1 and R_2 and how to obtain the asymptotic distribution of T_{opt}

Let $\rho = Corr(R_1, R_2) = Cov(R_1, R_2)$. First, we need to calculate $\hat{\rho}$ using the observed data:

$$\begin{aligned} Cov(S_1, S_2) &= E(S_1 S_2) - E(S_1)E(S_2) = E(S_1 S_2) = E[\sum_{i=1}^{m}(X_i - E(X_i))\sum_{i=1}^{m}|X_i - E(X_i)|] \\ &= \sum_{i=1}^{m}\sum_{j=1, j\neq i}^{m} E[(X_i - E(X_i))|X_j - E(X_j)|] + \sum_{i=1}^{m} E[(X_i - E(X_i))|X_i - E(X_i)|] \end{aligned}$$

If X_i and X_j are independent and since $E(S_1) = 0$,

$$\begin{aligned} Cov(S_1, S_2) &= \sum_{i=1}^{m}\sum_{j=1, j\neq i}^{m} E[(X_i - E(X_i))]E[|X_j - E(X_j)|] + \sum_{i=1}^{m} E[(X_i - E(X_i))|X_i - E(X_i)|] \\ &= \sum_{i=1}^{m} E[(X_i - E(X_i))|X_i - E(X_i)|] \\ \widehat{Cov}(S_1, S_2) &= \sum_{i=1}^{m} [(-4\hat{p_i}^2) * (1-\hat{p_i})^2 + (1-2\hat{p_i})^2 * (2\hat{p_i}(1-\hat{p_i})) + (2-2\hat{p_i})^2 * \hat{p_i}^2] \\ &= \sum_{i=1}^{m} 2\hat{p_i}(1-\hat{p_i})(1-2\hat{p_i})^2 \end{aligned}$$

Then,

$$\begin{aligned} \rho = Corr(R_1, R_2) &= Cov(R_1, R_2) = Cov(\frac{S_1 - E(S_1)}{\sqrt{Var(S_1)}}, \frac{S_2 - E(S_2)}{\sqrt{Var(S_2)}}) \\ &= \frac{1}{\sqrt{Var(S_1)}\sqrt{Var(S_2)}} Cov(S_1, S_2) \\ \hat{\rho} &= \frac{\sum_{i=1}^{m} 2\hat{p_i}(1-\hat{p_i})(1-2\hat{p_i})^2}{\sqrt{\sum_{i=1}^{m}(2\hat{p_i}(1-\hat{p_i}))\sum_{i=1}^{m}(2\hat{p_i}(1-\hat{p_i})(1-8\hat{p_i}(1-\hat{p_i})^3))}} \end{aligned}$$

If there are LD between the SNPs,

$$
\begin{aligned}
\widehat{Cov}(S_1, S_2) &= \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} \left[\sum_{X_i=0}^{2} \sum_{X_j=0}^{2} (X_i - E(X_i))|X_j - E(X_j)|\widehat{P_{i,j}}(X_i, X_j) \right] \\
&\quad + \sum_{i=1}^{m} E[(X_i - E(X_i))|X_i - E(X_i)|] \\
&= \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} \left[\sum_{X_i=0}^{2} \sum_{X_j=0}^{2} (X_i - E(X_i))|X_j - E(X_j)|\widehat{P_{i,j}}(X_i, X_j) \right] \\
&\quad + \sum_{i=1}^{m} 2\hat{p_i}(1-\hat{p_i})(1-2\hat{p_i})^2 \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\sum_{X_i=0}^{2} \sum_{X_j=0}^{2} (X_i - 2\,\hat{p_i})|X_j - 2\,\hat{p_j}|\widehat{P_{i,j}}(X_i, X_j) \right] + \sum_{i=1}^{m} 2\hat{p_i}(1-\hat{p_i})(1-2\hat{p_i})^2
\end{aligned}
$$

where $\widehat{P_{i,j}}(X_i, X_j)$ and $\hat{p_i}$ needs to be estimated using the dataset.

Then,

$$
\begin{aligned}
\hat{\rho} &= \frac{\widehat{Cov}(S_1, S_2)}{\sqrt{\widehat{Var}(S_1)}\sqrt{\widehat{Var}(S_2)}} \\
&= \frac{\widehat{Cov}(S_1, S_2)}{\sqrt{\sum_{i=1}^{m} \widehat{Var}(\Delta X_i) + \sum_{i=1}^{m} \sum_{j\neq i} \widehat{Cov}(\Delta X_i, \Delta X_j)}\sqrt{\sum_{i=1}^{m} \widehat{Var}(|\Delta X_i|) + \sum_{i=1}^{m} \sum_{j\neq i} \widehat{Cov}(|\Delta X_i|, |\Delta X_j|)}}
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{Cov}(\Delta X_i, \Delta X_j) &= \sum_{X_i=0}^{2} \sum_{X_j=0}^{2} (X_i - 2\,\hat{p_i})(X_j - 2\,\hat{p_j})\widehat{P_{i,j}}(X_i, X_j) \\
\widehat{Cov}(|\Delta X_i|, |\Delta X_j|) &= \sum_{X_i=0}^{2} \sum_{X_j=0}^{2} |X_i - 2\,\hat{p_i}||X_j - 2\,\hat{p_j}|\widehat{P_{i,j}}(X_i, X_j) \\
\widehat{Var}(\Delta X_i) &= \sum_{i=1}^{m} (2\,\hat{p_i}(1-\hat{p_i})) \\
\widehat{Var}(|\Delta X_i|) &= \sum_{i=1}^{m} (2\,\hat{p_i}(1-\hat{p_i})(1-8\,\hat{p_i}(1-\hat{p_i})^3))
\end{aligned}
$$

Therefore, since both R_1 and R_2 are also asymptotically distributed as $N(0,1)$, we could simulate the asymptotic distribution of T_{opt} for a particular dataset in the following steps. Suppose we use sample size $n = 100000$.

1. Suppose there are two random variables Z and W following two independent standard normal distributions. We generate n data points for the two random variables.

2. Create a new variable $U = \rho Z + \sqrt{(1-\rho^2)}$ W, then U also has a standard normal distribution. Also,

$$Corr(Z,U) = Cov(Z,U) = Cov(Z, \rho Z + \sqrt{(1-\rho^2)}W) = Cov(Z, \rho Z) = \rho$$

3. We generate the data points of U using the data points generated for Z and W.

4. We generate the data points of T using $T = max(Z^2, U^2)$.

Then, the simulated distribution of T is just the asymptotic distribution of T_{opt} for the dataset.

For the simulation of the asymptotic distribution of the test statistic in the simulation studies, we found that the asymptotic distribution does not vary much with the estimation of the correlation coefficient in the presence of LD between the markers and under the assumption of no LD between the markers. Thus, we decided to use the estimate of the correlation of R_1 and R_2 under the assumption of no LD between the markers to save computational time. From the results section we can see that this approximation is acceptable in terms of power and type I error.

A.3 Estimated Power of T_1 and T_2

Suppose the allele frequencies of the study population are $p_i, i = 1, ...m$ for the m SNPs, and the allele frequencies for the subpopulation that the outlier is from are pd_i where $pd_i = p_i + \delta_i$ for $i = 1, ...m$ markers, and assume that δ_i are independently and identically distributed, then we can analytically obtain the power of the statistics T_1 and T_2 and look at how the power changes as a function of p_i.

Suppose the outlier is from the subpopulation d with allele frequencies pd_i for $i = 1, ..., m$ markers, then the expected values of R_1 and R_2 can be obtained as the following:

$$E_d(R_1) = E_d\left(\frac{\sum_{i=1}^{m}(X_i - 2p_i)}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i))}}\right) = \frac{E_d\left[\sum_{i=1}^{m}(X_i - 2p_i)\right]}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i))}} = \frac{\sum_{i=1}^{m} E_d(X_i - 2p_i)}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i))}}$$

$$= \frac{\sum_{i=1}^{m} -2\,p_i\,(1-p_i-\delta_i)^2 + 2\,(1-2\,p_i)\,(p_i+\delta_i)\,(1-p_i-\delta_i) + (2-2\,p_i)\,(p_i+\delta_i)^2}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i))}}$$

$$= \frac{\sqrt{2}\sum_{i=1}^{m}\delta_i}{\sqrt{\sum_{i=1}^{m} p_i\,(1-p_i)}}$$

Thus, if there is a one-direction difference in the allele frequencies, i.e. $E(\delta_i) \neq 0$, then $T_1 = R_1^2$ has power to detect the population substructure in the dataset. Figure A.1 shows how the expected value of R_1 changes as a function of δ_i and average p_i, which indicates the change in power as a function of δ_i and p_i. As expected, the power is much larger for sequence data.

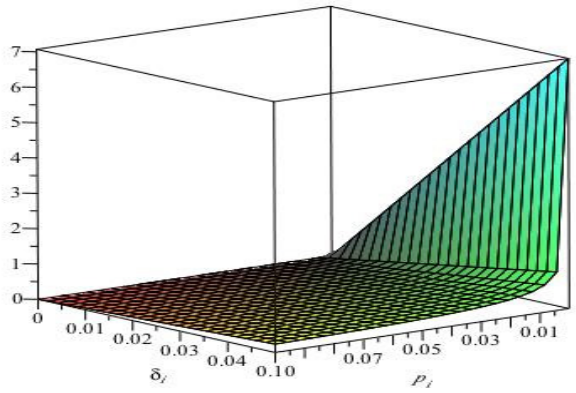

Figure A.1: The average power of R_1 as a function of δ_i and p_i.

Similarly we can obtain the expected value of R_2 for the outlier as the following: $E_d(R_2) =$

$$E_d\left(\frac{\left[\sum_{i=1}^{m}|X_i-2p_i|-(4p_i(1-p_i)^2)\right]}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))}}\right)$$
$$=\frac{E_d\left[\sum_{i=1}^{m}|X_i-2p_i|-(4p_i(1-p_i)^2)\right]}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))}}$$
$$=\frac{\sum_{i=1}^{m}\left[E_d(|X_i-2p_i|)-(4p_i(1-p_i)^2)\right]}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))}}$$
$$=\frac{[\sum_{i=1}^{m}\left[2\,p_i(1-p_i-\delta_i)^2+2\,(1-2\,p_i)(p_i+\delta_i)(1-p_i-\delta_i)+(2-2\,p_i)(p_i+\delta_i)^2-(4p_i(1-p_i))\right]]}{\sqrt{\sum_{i=1}^{m}(2p_i(1-p_i)(1-8p_i(1-p_i)^3))}}$$
$$=\frac{\sum_{i=1}^{m}\left[4\,p_i\,\delta_i^2+(8\,p_i^2-8\,p_i+2)\,\delta_i\right]}{\sqrt{\sum_{i=1}^{m}(2\,p_i(1-p_i)(1-8\,p_i\,(1-p_i)^3))}}$$
$$=\frac{\sum_{i=1}^{m}4\,p_i\,\delta_i^2}{\sqrt{\sum_{i=1}^{m}(2\,p_i(1-p_i)(1-8\,p_i\,(1-p_i)^3))}}+\frac{\sum_{i=1}^{m}(8\,p_i^2-8\,p_i+2)\,\delta_i}{\sqrt{\sum_{i=1}^{m}(2\,p_i(1-p_i)(1-8\,p_i\,(1-p_i)^3))}}$$

This formula shows clearly that even if there is two-direction difference between the allele frequencies in the two populations, i.e. $E(\delta_i) = 0$, R_2 still has power since it includes the δ_i^2 term which is not expected to be 0. Figure A.2 shows how the expected values of R_2 changes as a function of δ_i and p_i, and again if $E(\delta_i) \neq 0$, it shows that the power increases rapidly as the average allele frequency becomes much smaller than 0.01. If $E(\delta_i) = 0$, the expected test statistic changes with the term in front of δ_i^2, and actually increases with p_i as shown in Figure A.3.

We have also considered test statistics with higher moments of $\Delta X_i = X_i - 2\,p_i$, and looked at how the power changes as a function of the moment and p_i. Let

$$R_q = \frac{[\sum_{i=1}^{m}(X_i - 2\,p_i)^q - E\,[(X_i - 2\,p_i)^q]}{\sqrt{\sum_{i=1}^{m} Var((X_i - 2\,p_i)^q)}}$$

Then the expected value of R_q for the outlier is:

$$E_d(R_q) = E_d\left[\frac{[\sum_{i=1}^{m}(X_i-2\,p_i)^q-E[(X_i-2\,p_i)^q]}{\sqrt{\sum_{i=1}^{m}Var((X_i-2\,p_i)^q)}}\right] = \frac{[\sum_{i=1}^{m}E_d[(X_i-2\,p_i)^q]-E[(X_i-2\,p_i)^q]}{\sqrt{\sum_{i=1}^{m}Var((X_i-2\,p_i)^q)}}$$

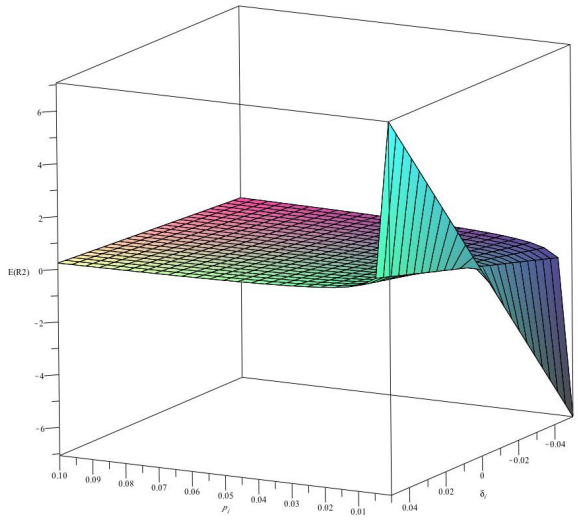

Figure A.2: The average power of R_2 as a function of δ_i and p_i.

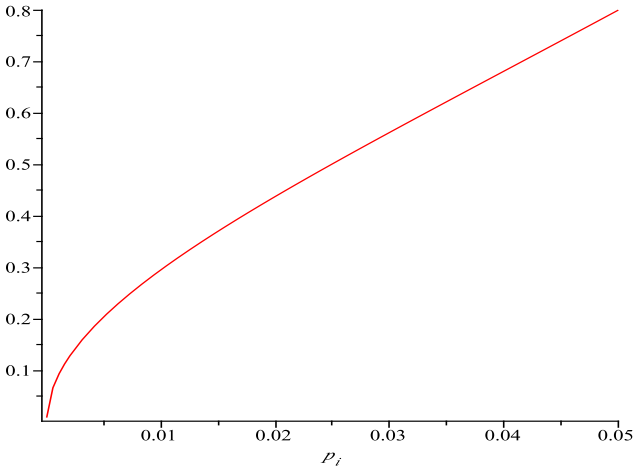

Figure A.3: The average power of R_2 as a function of p_i if $E(\delta_i) = 0$.

$$= \frac{\left[\sum_{i=1}^{m}(-2\,(1-2\,p_i)^q p_i + 2\,(1-2\,p_i)^q(1-p_i) + (-2\,p_i)^q(-2+2\,p_i) + 2\,(2-2\,p_i)^q p_i)\delta_i\right]}{\sqrt{\sum_{i=1}^{m} Var((X_i - 2\,p_i)^q)}}$$

$$+ \frac{\left[\sum_{i=1}^{m}((-2\,p_i)^q - 2\,(1-2\,p_i)^q + (2-2\,p_i)^q)\delta_i{}^2\right]}{\sqrt{\sum_{i=1}^{m} Var((X_i - 2\,p_i)^q)}}$$

where

$$Var\left[(X_i - 2\,p_i)^q)\right] = (-2\,p_i)^q\,(1-p_i)^2 + 2\,\,(1-2\,p_i)^q\,p_i\,(1-p_i) + (2-2\,p_i)^q\,p_i{}^2$$

If $E(\delta_i) = 0$, the term in front of δ_i^2 determines the power, thus we plot this term as a function of q and p_i and we observe that for different MAF level, with q greater than a threshold approximately, the expected statistic does not increase as rapidly as for q less or equal to that threshold. For a fixed moment q, the expected value of $T_q = R_q^2$ increases very fast as the allele frequency decreases.

From Figure A.4, it seems we should use greater moment to achieve better power, however, in our simulation to compare the power of the test statistics based on $\sum_{i=1}^{m} |\Delta X_i|$ and $\sum_{i=1}^{m} (\Delta X_i)^2$, we obtain better power with $\sum_{i=1}^{m} |\Delta X_i|$. We believe that this is due to the fact that most of the power for the scores based on higher moments is from the subjects with homozygous rare allele type. However, the number of subjects with homozygous rare alleles is very limited in real datasets. Also, there is not much contribution from subjects with heterozygous genotype for scores based on higher moments. Thus, with limited number of subjects in the rare variant dataset as in most real-world scenarios, the power of $\sum_{i=1}^{m} |\Delta X_i|$ is higher. Thus, we decide to use $\sum_{i=1}^{m} |\Delta X_i|$ rather than $\sum_{i=1}^{m} (\Delta X_i)^2$, or even higher moments.

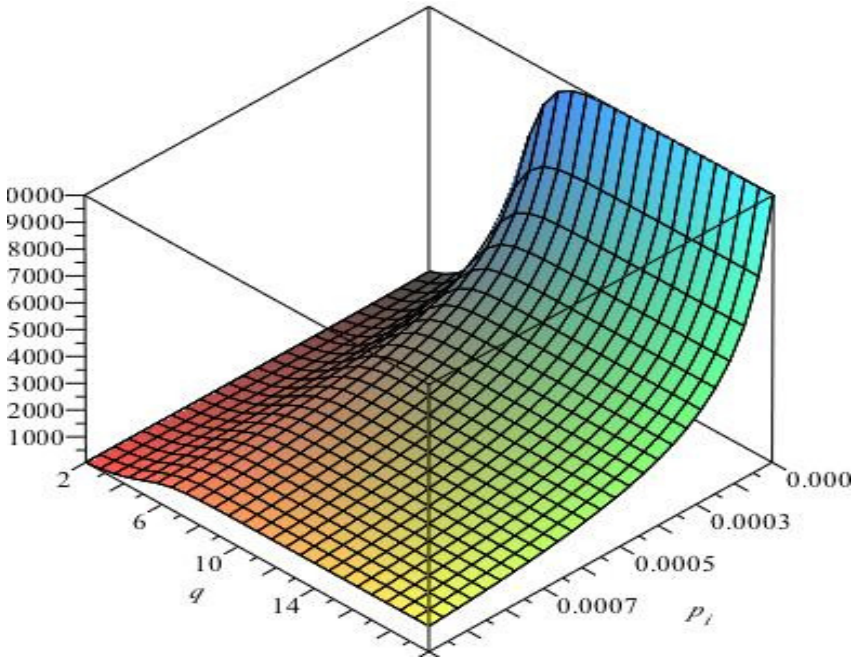

Figure A.4: The average power of R_2 as a function of p_i if $E(\delta_i) = 0$.

Appendix B

Sensitivity analysis of the Bin test

B.1 Sensitivity analysis

There are two variables to be investigated in the sensitivity analysis: the p-value cut-off percentile – P, the pre-defined proportion of neighboring variants for calculating the distances – R.

In our implementation, there are M permutations in each permutation set. More specifically, this means that a pre-defined M number of permutations are first generated to calculate the p-value of clustering, then another set of M permutations are obtained to test this p-value (two-sided Binomial test). If the p-value is rejected using the second set of M permutations, a new set of M permutations are generated in addition to the previous ones to obtain the updated p-value and another set of M permutations are used to test this p-value, this continues until the p-value is not rejected by the last set of M permutations. This pre-defined number of permutations M are set to be 2000 in the sensitivity analysis of the power and type I error. If insufficient number of permutations are used, the type I error would be inflated.

We are interested in the effect of the p-value cut-off P, and the range threshold R for the number of neighboring variants for the calculation of distances on the power and type I error rate of the test.

B.1.1 Power

The same simulation model and parameters are used here as scenario 1 described in the simulation section. The effect sizes are the same as for Effect 1 in Table 3.1. Three levels of the p-value cut-off are considered: 0.5%, 1%, and 5%. Three levels of the threshold for the number of neighboring variants used in the calculation

Table B.1: The power of the Bin test with different quantiles for the p-value cut-off threshold P and the number of neighboring SNP for calculating the distances.

Power	0.5%	1%	5%
0.01%	0.77	0.64	0.45
0.05%	0.89	0.79	0.60
0.1%	0.91	0.87	0.77

2000 permutation set is used in each replicate and 100 replicates are used in each case. The columns correspond to three quantiles for p-value cut-off. The rows correspond to three quantiles for the number of neighboring variants to be used for calculating distances.

of distances are considered: 0.01%, 0.05%, and 0.1%. Note that the values are the quantiles among all the SNPs rather than the absolute value. The results are shown in Table B.1 and we observe that as the p-value cut-off P decreases (a smaller number of variants would be included), the power of the test increases. This is expected because the causal variants with more significant p-value would give more information and less noise in the calculation. However, if only the top p-values are included, causal variants with relatively small p-value in proximity to other causal variants would be ignored, which in turn would decrease the power. Also, as the threshold for the number of neighboring variants R increases, the power of the test increases as expected, but more computation and time would be needed.

B.1.2 Type I Error

With 2000 permutation sets, the type I error is mostly well-maintained, as in Table B.2. However, with a smaller number of permutations used, we do observe an increase in

type I error as the p-value cut-off P decreases and as the threshold R for the number of neighboring SNPs increases (data not shown). Thus, from our simulations, it is recommended to use a larger number of permutations to obtain the p-value for a small p-value cut-off P and large threshold R for the number of neighboring SNPs.

Table B.2: The type I error of the Bin test with different p-value cut-off threshold P and different range R for the neighboring SNP for calculating the distances.

Type I Error	0.5%	1%	5%
0.01%	0.04	0.04	0.02
0.05%	0.06	0.04	0.03
0.1%	0.02	0.04	0.06

2000 permutation set is used in each replicate and 100 replicates are used. The columns correspond to three quantiles for p-value cut-off. The rows correspond to three quantiles for the number of neighboring variants to be used for calculating distances.

B.2 Binary search for the associated region

One advantage of our test is that it could detect a loci or a gene that is clustered with causal variants in a large region, thus it could be used in a binary search for the associated region with the phenotype. The idea is to test each chromosome separately first, then do a binary search for the chromosome with significant p-values. A tentative procedure is:

1) Define the minimum size of a loci to be tested. Check if the first and the second half of the region are greater than this minimum size. If not, this region should be returned as the associated region with the phenotype. Otherwise, the first and second half of the region are tested separately.

2) If none of the half of the region is significant, the middle region should be tested, go to step 3); If any of the half of the region is significant, repeat step 1) for the significant region.

3) Define a maximum size of a loci to be tested and a step size. If the middle region with the predefined minimum size is tested to be significant, report this region. Otherwise,

the middle region is increased at the two end with the step size and tested again until either significant result is obtained or the maximum size is reached. If no significant region found even for the maximum size, report the last region with a significant result. If significant result if obtained, report this middle region.